

**MEDICAL STATISTICS
and COMPUTER
EXPERIMENTS**

2nd Edition

THE
JOURNAL OF
THE
ROYAL ANTHROPOLOGICAL INSTITUTE
OF GREAT BRITAIN AND IRELAND
PUBLISHED BY THE
CAMBRIDGE UNIVERSITY PRESS

Volume 100, Part 1, 2000



MEDICAL STATISTICS and COMPUTER EXPERIMENTS

2nd Edition

Editor

Ji-Qian Fang

Sun Yat-Sen University, P R China

with

Yongyong Xu

*Fourth Military Medical
University, P R China*

Songlin Yu

*Huazhong University of Science
and Technology, P R China*

 **World Scientific**

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data

Medical statistics and computer experiments / [edited by] Ji-Qian Fang, Sun Yat-Sen University, China. -- 2nd edition.

pages cm

Includes bibliographical references and index.

ISBN 978-981-4566-77-3 (hardcover : alk. paper)

1. Medical statistics. 2. Medicine--Research--Statistical methods. 3. Medicine--Research--Data processing. 4. Medicine--Research--Computer simulation. I. Fang, Ji-Qian, 1939-- editor.

R853.S7.M43 2014

610.72'7--dc23

2013041200

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

Copyright © 2014 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

Typeset by Stallion Press

Email: enquiries@stallionpress.com

Printed in Singapore by B & Jo Enterprise Pte Ltd

Preface to the Second Edition

Medical Statistics and Computer Experiments (Chinese version) was recommended as a textbook for graduate students by the Ministry of Education, People's Republic of China. There have been four editions of Chinese version published by the Shanghai Science and Technology Press in 1997, 2001, 2006 and 2012. The first English version, whose content was based on the third Chinese edition was published by World Scientific in 2005.

The idea of updating the English version was first raised by the publisher. The publisher has been forwarding the reviews of the first edition to me timely. I have been inspired whenever I read those reviews. Meanwhile, due to the increasing demand of the third edition of the Chinese version, the Shanghai Science and Technology Press was also looking forward to a new edition. Therefore, we decided to update both the Chinese and English versions simultaneously, and as a result, the fourth edition of Chinese version had been published in June 2012.

In order to have a new edition, I decided to update at least 30% of the content. For chapters which were essential for the compulsory and elective courses of *Medical Statistics*, and basic for the healthcare related researches, we have maintained the fundamental content by "getting rid of the stale and taking in the fresh". For example, "histogram" is essential, but it is hard to decide which is superior to the other when two histograms are almost the same. In this case, after "deleting some superfluous", we have included the *cumulative frequency diagram*, which clearly presents two ladder-shaped curves, with one sitting above the other.

Furthermore, the core of ANOVA should be the concept on decomposition of the total sum of squared deviations according to the sources of variation, while the convenience in calculation takes a second role. In the new edition, instead of the formulas for hand calculating, we emphasize the

meaning of the formulas for all kinds of sum of squares, and the fundamental difference among various designs. In addition, based on the formula of sum of squared errors, the residual analyses are uniformly applied to examine the preconditions of all kinds of ANOVA.

Moreover, to serve the demands of new medical researches, some of the original content has been enriched with new concepts and knowledge. For example, to meet the increasing needs of large-scale clinical trials, the non-inferiority test, equivalence test and interim analysis, which are receiving more and more attention, have been introduced and explained in an easy to understand way. Again, as the data frequently encountered in practice do not typically follow a presumed distribution, the approach of permutation test has been provided, which is gradually becoming popular as a supplement to the routine tests.

Finally, as modern clinical research pays equal attention to both head-to-head comparative trials to investigate clinical effects in the real world on the macroscopic aspect, and massive genetic and molecular information from the microscopic perspective, two chapters illustrating comparative effectiveness research and design and analysis of bioinformatics have been newly added, in addition to the improvement of the previous chapters on statistical methods of scale research and genetic statistics.

The first Chinese edition recommended the brand-new teaching method of "*Computer experiments on statistics*". Since then, this method has been acknowledged by peers from home and abroad and found its way into more and more schools. In this edition, some new experiments related to the enriched text have been added. For example, in chapters introducing random permutation test, propensity matching and differential item functioning, the corresponding SAS programs have been provided, which are helpful in dealing with massive genetic data, observational data and multi-center scale data. As before, all SAS programs and some SPSS programs are shown and are also available at

<http://www.worldscientific.com/r/8981-suppl>, please register/sign in at the website.

The new edition was written mainly during summer vacation. Professor Song-Lin Yu worked hard in Wuhan, known as one of the "four furnaces" in China, and sometimes even replied my email discussing the text at midnight. Professor Yong-Yong Xu performed "major surgery" on chapters

of ANOVA and sequential analysis, dramatically improving the layout. Professor Yuan-Tao Hao standardized the content of the appendices, and provided template at the same time. Led by director Jin-Xin Zhang and Professor Jing Gu, a group of postgraduate students majoring in medical statistics and epidemiology in the Sun Yat-Sen University scrutinized the text and appendices as readers to find any errors or defects. With profound knowledge on the subject, Professor Dong Yi wrote a new chapter on Bioinformatics based on recent literatures.

Many colleagues such as Prof. Chang-Sheng Chen, Dr. Dan-Hong Liu, Prof. Yi Wan and Dr. Chun Hao creatively took part in the arduous revision; and as they did before, Shao-Min Wu, Fang-Fang Zeng, and Shu-Min Zhu were helpful in graphing, typing and recording the appendices. Here, I would like to express my sincere thanks to all the above-mentioned and unmentioned colleagues and friends for their great contribution.

With full cooperation of the professors mentioned above, this edition has made considerable progress in both the content and form. Due to the constraint of time and energy, however, new errors and even mistakes might appear inevitably. Criticism and suggestions are always welcome so that I will be encouraged to continuously revise in next print or edition.

Ji-Qian Fang
April 2014



Introduction

This volume consists of three parts: Part I has 11 chapters on basic concepts of statistics, Part II includes ten chapters on multi-variate statistics and Part III includes 12 chapters on design and analysis for medical research. This volume uses basic concepts and commonly used methods on design and analysis in medical statistics, incorporating the operation of statistical package SAS and 100 computer experiments for the important statistical phenomena related to each chapter. Each chapter concludes with a section on "Practice and Experiments". All necessary data including reference answers for exercises, SAS programs for all computer experiments and part of the examples, data documents for 12 medical researches are available at <http://www.worldscientific.com/r/8981-suppl>, please register/sign in at the website. Part I of this volume can be used for a required course with 100 teaching hours or so; the last two parts can be used for a course with 50 teaching hours for each. This volume can also be used as a reference book for related courses in life science, agriculture and forestry.



Contents

Preface to the Second Edition	v
Introduction	ix
About the Editors	xxi

Part I Basic Concepts 1

Chapter 1. Descriptive Statistics 3

1.1 Variables and Data	3
1.2 Frequency Table and Histogram	6
1.3 Measurement for Average Level of a Sample	13
1.4 Measurement for Variation of a Sample	20
1.5 Relative Measures and Standardization Approaches	22
1.6 Frequently Used Graphs in Statistics	28
1.7 Computerized Experiments	35
1.8 Practice and Experiments	41

Chapter 2. Probability and Distribution 45

2.1 Explanation of Probability and Related Concepts	45
2.2 Distributional Characters of Random Variables	49
2.3 Binomial Distribution	53
2.4 Poisson Distribution	60
2.5 Normal Distribution	63
2.6 Computerized Experiments	71
2.7 Practice and Experiments	74

Chapter 3. Sampling Error and Confidence Interval	77
3.1 The Distribution of Sample Mean	77
3.2 t Distribution	83
3.3 The Confidence Interval for Population Mean of a Normal Distribution	85
3.4 Four Confidence Intervals for Probability and the Difference between Two Probabilities	87
3.5 The Sample Size for Estimation of Confidence Interval	88
3.6 Computerized Experiments	90
3.7 Practice and Experiments	93
Chapter 4. Hypothesis Testing for Continuous Variables	95
4.1 Specific Logic and Main Steps of Hypothesis Testing	95
4.2 The t Test for One Group of Data under Completely Randomized Design	99
4.3 The t Test for Data under Randomized Paired Design	101
4.4 The Tests for Comparing Two Means Based on Two Groups of Data under Completely Randomized Design	103
4.5 The F -Test for Equal Variances of Two Groups of Data under Completely Randomized Design	107
4.6 Test for Normality	110
4.7 The Z -Test for the Parameters of Binomial Distribution and Poisson Distribution (Large Sample)	112
4.8 Computerized Experiments	121
4.9 Practice and Experiments	125
Chapter 5. Chi-Square Test for Categorical Variable	131
5.1 Chi-Square Distribution and Pearson's Goodness-of-Fit Test	131
5.2 The χ^2 Test for Comparison between Two Independent Sample Proportions	134
5.3 The χ^2 Tests for Binary Variable under a Paired Design	142
5.4 The χ^2 Test for $R \times C$ Contingency Table	148
5.5 The χ^2 Test for Confirming a Supposed Distribution	153
5.6 Hypothesis Testing for Two Standardized Rates	155
5.7 Fisher's Exact Test for 2×2 Table	159

5.8	Computerized Experiments	163
5.9	Practice and Experiments	166
Chapter 6. Further Discussion on Hypothesis Test		171
6.1	Two Types of Error and Power	172
6.2	The Four Elements Affecting the Power	174
6.3	The Quantitative Relation between Power and the Four Elements	177
6.4	Estimation of Sample Size for the Tests in Common Use . . .	182
6.5	Non-Inferiority Test and Equivalence Test	185
6.6	Permutation Test	189
6.7	Computerized Experiments	192
6.8	Practice and Experiments	194
Chapter 7. Single-Factor Analysis of Variance		197
7.1	One-Way Analysis of Variance: Completely Random Design	197
7.2	Two-Way Analysis of Variance: Randomized Complete-Block Design	215
7.3	Three-Way Analysis of Variance: The Latin-Square Design	221
7.4	Computerized Experiments	229
7.5	Practice and Experiments	233
Chapter 8. Nonparametric Test Based on Ranks		237
8.1	Wilcoxon's Signed Rank Test	238
8.2	Wilcoxon's Rank-Sum Test for Comparing the Locations of Two Distributions	242
8.3	Hypothesis Testing for the Locations of More Than Two Populations	248
8.4	Computerized Experiments	257
8.5	Practice and Experiments	259
Chapter 9. Simple Linear Correlation		263
9.1	Concept of Correlation	263
9.2	Correlation Coefficient	266
9.3	Inference on Correlation Coefficient	269

9.4	Rank Correlation	272
9.5	Caution in Analysis of Linear Correlation	275
9.6	Computerized Experiments	277
9.7	Practice and Experiments	278

Chapter 10. Simple Linear Regression **281**

10.1	Statistical Description of Linear Regression	281
10.2	Statistical Inference on Regression	284
10.3	Applications of Linear Regression and the Pre-requisites	292
10.4	On the Basic Assumptions and Analysis of Residuals	299
10.5	Non-linear Regression	301
10.6	Computerized Experiments	309
10.7	Practice and Experiments	313

Chapter 11. Statistical Principles for Design of Interventional Study **317**

11.1	The Essential Concepts of Design	318
11.2	Statistical Principle in Clinical Trials	323
11.3	Randomization Techniques	332
11.4	Randomized Controlled Trial	336
11.5	Comments on Some Medical Examples	342
11.6	Computerized Experiments	345
11.7	Practice and Experiments	347

Part II Multi-variate Statistics **349**

Chapter 12. Multiple Regression and Correlation **351**

12.1	Basic Procedure of Multiple Regression	351
12.2	Multiple Correlation	357
12.3	Selection of Independent Variables	361
12.4	Further Topics in Multiple Regression	365
12.5	Path Analysis	373
12.6	Computerized Experiments	378
12.7	Practice and Experiments	380

Chapter 13. Measures of Multi-variate Data and Multi-variate Analysis of Variance	383
13.1 Multi-variate Statistical Description	383
13.2 Comparison between Two Mean Vectors — Hotelling's T^2 Test	388
13.3 Comparisons among Several Multi-variate Means—Multi-variate Analysis of Variance	392
13.4 Computerized Experiments	397
13.5 Practice and Experiments	400
Chapter 14. Discriminant Analysis	403
14.1 Basic Ideas of Discriminant Analysis	403
14.2 Fisher's Discriminant Analysis	405
14.3 Bayesian Discriminant Analysis	407
14.4 Stepwise Discriminant Function	411
14.5 Decision Tree	413
14.6 Retrospective and Prospective Validation	422
14.7 Considerations in Applications	424
14.8 Computerized Experiments	426
14.9 Practice and Experiments	428
Chapter 15. Logistic Regression	431
15.1 Logistic Regression Model	431
15.2 Conditional Logistic Regression	445
15.3 Multinomial Logistic Regression Model	447
15.4 Two-Level Logistic Mixed Effects Regression Model	453
15.5 Application of Logistic Regression	456
15.6 Computerized Experiments	458
15.7 Practice and Experiments	462
Chapter 16. Cluster Analysis	465
16.1 The Meaning of Clustering	465
16.2 Hierarchical Cluster	467
16.3 Fast Cluster	471
16.4 Variable Cluster	473

16.5	Computerized Experiments	474
16.6	Practice and Experiments	477
Chapter 17. Principal Component Analysis		479
17.1	The Basic Concepts of Principal Component Analysis	479
17.2	Computation and Interpretation of Principal Components	483
17.3	Principal Component Analysis in Regression	487
17.4	Computerized Experiments	490
17.5	Practice and Experiments	493
Chapter 18. Factor Analysis		497
18.1	Factor Model	497
18.2	Derivation of Factors	498
18.3	Factor Pattern Plot and Factor Rotation	502
18.4	Factor Score and Application of Factor Patterns	507
18.5	Confirmatory Factor Analysis	508
18.6	Computerized Experiments	514
18.7	Practice and Experiments	515
Chapter 19. Canonical Correlation and Correspondence Analysis		517
19.1	Canonical Correlation	517
19.2	Correspondence Analysis	528
19.3	Canonical Discriminant Analysis	535
19.4	Computerized Experiments	538
19.5	Practice and Experiments	539
Chapter 20. Survival Analysis		541
20.1	The Basic Concept of Survival Analysis	542
20.2	The Product-Limit Method for One Group of Survival Data	543
20.3	The Log-Rank Test and Breslow Test for Comparing Two Survival Data Sets	547
20.4	The Cox Regression	552
20.5	Computerized Experiments	558
20.6	Practice and Experiments	558

Chapter 21. Log-Linear Model for Contingency Table and Poisson Regression	561
21.1 Log-Linear Models for Contingency Table	561
21.2 Poisson Regression	574
21.3 Computerized Experiments	578
21.4 Practice and Experiments	581
 Part III Design and Analysis for Medical Research	 583
Chapter 22. Multi-Factor Analysis of Variance	585
22.1 Factorial Experiments and Analysis of Variance	585
22.2 Split-Plot Designs and Analysis of Variance	593
22.3 Cross-Over Design and Analysis of Variance	604
22.4 Computerized Experiments	609
22.5 Practice and Experiments	611
 Chapter 23. Analysis of Repeated Continuous-Type Measurements	 615
23.1 Examples of Repeated Measurements	615
23.2 Imperfect Analysis and its Origins	618
23.3 Approach with Summary Measures	619
23.4 Analysis of Variance for Repeated Measurements	620
23.5 Computerized Experiments	629
23.6 Practice and Experiments	631
 Chapter 24. Design and Analysis of Cross-Sectional Studies	 633
24.1 Design of the Study	633
24.2 Sampling Methods and Estimation of Population Parameters	634
24.3 Estimation of Sample Size	642
24.4 The Current Life Table	648
24.5 Computerized Experiments	657
24.6 Practice and Experiments	659
 Chapter 25. Design and Analysis of Prospective Studies	 661
25.1 Study Design	661
25.2 Measures of Disease Occurrence	663

25.3	Analysis of Data from Prospective Studies	671
25.4	Computerized Experiments	684
25.5	Practice and Experiments	692
Chapter 26.	Designs and Analysis of Case-Control Studies	693
26.1	Designs of Case-Control Studies	693
26.2	Analysis of Data from Design for Group Comparison	700
26.3	Analysis of Matched Data	710
26.4	Computerized Experiments	717
26.5	Practice and Experiments	719
Chapter 27.	Design and Analysis of Diagnostic and Screening Tests	721
27.1	Design and Data Layout	721
27.2	Measures Frequently Used in Diagnostic Test	721
27.3	Analysis of ROC Curve	727
27.4	Decision Making on Diagnostic and Screening Test	735
27.5	Computerized Experiments	739
27.6	Practice and Experiments	742
Chapter 28.	Design and Analysis of Sequential Experiments	747
28.1	Introduction	747
28.2	Design and Analysis of Sequential Trials	748
28.3	Group Sequential Schemes	755
28.4	Computerized Experiments	763
28.5	Practice and Experiments	764
Chapter 29.	Systematic Review of Medical Research and Meta-Analysis	767
29.1	Basic Notions	767
29.2	Statistical Methods Commonly Used in Meta-Analysis	773
29.3	Notes	784
29.4	Computerized Experiments	787
29.5	Practice and Experiments	790
Chapter 30.	Comparative Effectiveness Research	793
30.1	Background	793

30.2	Definitions	794
30.3	Examples	796
30.4	Features and Principles	799
30.5	Research Methods and Techniques	802
30.6	Steps of CER	819
30.7	Standards for Implementation and Report	822
30.8	Summary	824
30.9	Computerized Experiments	825

Chapter 31. Statistical Methods in Scale Development 831

31.1	Development of Scales	831
31.2	Adopting Scale with Foreign Language	835
31.3	The Concept and Evaluation of Validity and Reliability	840
31.4	Item Response Theory and Scale Evaluation	851
31.5	Computer Experiments	856
31.6	Exercises and Experiments	856

Chapter 32. Statistical Methods for Data from Genetic Epidemiological Study 859

32.1	Basic Concepts	859
32.2	Linkage Analysis	865
32.3	Genetic Association Analysis	871
32.4	Computerized Experiments	877
32.5	Practice and Experiments	879

Chapter 33. Statistical Methods in Bioinformatics 881

33.1	Sequence Alignment Methods	882
33.2	The Data Acquisition and Standardization of Gene Expression Patterns	885
33.3	Differentially Expressed Genes Screening	887
33.4	Cluster Analysis of Gene Expression	890
33.5	Analysis of Gene Regulatory Networks	900
33.6	Computerized Experiments	904
33.7	Summary	906
33.8	Practice and Experiment	908

Appendix I. Introduction to the Statistical Analysis System (SAS)*	
Appendix II. Statistical Tables	909
Appendix III. Datasets of Some Real Medical Examples	983
Appendix IV. Answers to Exercises*	
Appendix V. SAS Programs and Data*	

*Appendices I, IV and V are available at <http://www.worldscientific.com/r/8981-suppl>, please register/sign in at the website.

About the Editors



Professor **Ji-Qian Fang** was awarded “National Teaching Master” in 2009, and “Outstanding Contribution to Preventive Medicine” in 2010 by the Chinese Central Government. He is the leading professor in research and education of medical statistics in China. He was born in Shanghai 1939, earned his BS in 1961 from Department of Mathematics, Fudan University and Ph.D. in 1985 from the Program of Biostatistics, University of California at Berkeley.

His Ph.D. thesis studied multi-state survival analysis for life phenomena under the guidance of Professor Chin-Long Chiang. During 1985–1990, he was a Professor and Director, Department of Biostatistics and Biomathematics, Beijing Medical University. Since 1991, he has been the Director and Chair Professor, Department of Medical Statistics, Sun Yat-Sen University. Professor Fang was a visiting professor of University of Kent, UK in 1987 and Australian National University in 1990, as well as an adjunct professor of Chinese University of Hong Kong in 1993–2009. He is the secretary for the International Biometric Society and vice president of Chinese Association of Health Informatics.

Professor Fang has supervised many post-graduate students and post-doctor fellows in medical statistics. His research projects cover widely various fields, including “Stochastic Models of Life Phenomena”, “Gating Dynamics of Ion Channels”, “Statistical Methods for Data on Quality of Life”, “Health and Air Pollution”, “Analysis of DNA Finger Printing”, and

“Linkage Analysis between Complex Trait and Multiple Genes” etc. These projects were sponsored either by the National Foundation of China or by international organizations, such as World Health Organization and European Commission.

He is the chief editor of the textbook *Health Statistics* in Chinese (5th–7th edns.), *Medical Statistics and Computer Experiments* in English (1st and 2nd edns.) and the co-editor of the monograph *Advanced Medical Statistics* in Chinese and English respectively (1st and 2nd edns.).



Yongyong Xu, MSc and Ph.D. in medical statistics, he is the director of the Department of Health Statistics of the Fourth Military Medical University in Xi'an, Shaanxi province of China. Dr Xu has been a professor and taught medical statistics for undergraduate students, master students, Ph.D. students and medical professionals since 1988. He has edited four textbooks as chief editor in medical statistics for medical students and published for Chinese medical colleges.

His main research areas are statistical methods in medical research, clinical trials and statistical computing. In recent years, his area of research covers statistical methodology in analysis of health administrative data, with emphasis on the problems associated with growth standards for Chinese children and adolescent, case-mix, strategy for building Chinese health information standard framework, planning and improving the national health indicator framework and statistics. Dr Xu was selected as chairman of the Chinese Society of Health Information Standardization, honorary chairman of Shaanxi Society of Health Statistics, and chairman of PLA Association of Health Statistics. He was also awarded the prizes as the model of Ph.D. students' supervisor of Shaanxi province and the National Teaching Model.



Songlin, Yu, emeritus professor of medical statistics at Tongji Medical College, Huazhong University of Science and Technology in Wuhan, China. He received his medical doctor degree in Tongji Medical University in 1960 and he has over 44 years of experience teaching medical statistics, regression analysis, survival analysis, multivariate analysis and design of experiments to both undergraduate and graduate students. He taught time series analysis and econometrics for students of health statistical specialty. His several research programs granted by the Chinese government, National Foundation of Natural Sciences, and World Health Organization (WHO) have excellent successes and awards. He has also published a number of articles in statistical and applied field studies. Since early 1990, he was involved in health behavior research programs such as smoking control. He wrote three monographs: *"Statistical Methods for Field Studies in Medicine"*, *"Survival Analysis in Clinical Researches"* and *"Analysis of Repeated Measures Data"*. Together with Chinese colleagues, he participated in writing several textbooks in medical statistics for students of different levels. He was officially appointed as editor-in-chief responsible for writing and editing a textbook *"Medical Statistics"* in 2001. This book had been reprinted four times in 2002, 2003 and 2004. He, as coauthor, wrote a new book entitled *"R and Applications in Environmental Epidemiology"* (in press).

Professor Yu was the member of National Committee of New Drug Evaluation, Ministry of Public Health, China (1995–1998), the vice-president of Hubei Provincial Association of Health Statistics (1994–2005), the permanent member of Committee and vice-president of the National Association of Medical Statistics, Society of Preventive Medicine of China (1993–2002). He is also a member of editorial board and peer reviewer for many Chinese journals in health sciences and clinical medicine.



Part I

Basic Concepts



Chapter 1

Descriptive Statistics

Statistical analyses in practice usually include two parts: statistical description and statistical inference. Statistical description is a kind of fundamental work for statistical inference, which describes the feature of the sample. The main forms for description are tables (such as frequency table), plots (such as block plot, histogram) and numerical indices (such as mean, standard deviation).

1.1 Variables and Data

1.1.1 *Types of variables*

Variables are used to describe the properties of individuals in statistics. Different types of variables have different types of distributions and hence the statistical methods being used might be different. It is important to identify the types of variables before dealing with the data.

1.1.1.1 *Continuous variable*

They are the variables whose values can be obtained through measurement such as height, weight, blood pressure, pulse and blood count of the individuals. Limited by the precision of measurement, the variables such as height and weight can take some values of real number but not all indeed, and the variables such as pulse and blood count can take values of integral number only. However, for the convenience in theoretical study, they are regarded as continuous variables taking values in a continuous interval on the axis of real number. Sometimes, the observed values of such kind of variables are called measurement data.

1.1.1.2 Discrete variable

Some properties can only be described qualitatively with several mutually excluded categories, such as gender, occupation and effect of medicine (positive or negative). The variable for gender can only take a “value” either “male” or “female”; the variable of occupation may take a “value” among several categories (worker, farmer, salesman and soldier etc.). This kind of variables is called categorical variables or nominal variables.

Example 1.1 The variable for gender can be defined with a binary variable X .

$$X = \begin{cases} 0 & \text{Female,} \\ 1 & \text{Male.} \end{cases}$$

In general, the variables taking values in a set of countable numbers are called discrete variables. Binary variable is the simplest special case of it.

The number of individuals within a certain category is often counted, and it is called frequency so that the data of discrete variable is sometimes called count data.

Example 1.2 In the sample of 108 patients, there are 63 males and 45 females. If a binary variable X is defined for gender as in Example 1.1, the sum of X for the 108 patients is the number of males (63).

In general, the frequency of certain category is equivalent to the sum of a binary variable.

1.1.1.3 Ordinal variable

Some measurement can only result in a semi-quantitative outcome. For instance, $-$, \pm , $+$, $++$, $+++$ are quite often used to indicate different ranks in clinic. For some properties, there naturally exist ranks among different categories. For instance, cure, effective, un-effective and worse are used to describe the level of drug effect. An ordinal variable can be defined for this kind of properties taking values among 1, 2, 3, ... for rank, but not for the exact quantitative measurement.

The frequencies of ordinal variable is sometimes called ranked data.

Table 1.1 The post-treatment clinical records of 100 hypertension patients.

No.	Age (years)	Gender	Treatment	Systolic pressure (kPa)	Diastolic pressure (kPa)	ECG	Effectiveness
1	37	Male	Drug A	18.67	11.47	Normal	Prominent
2	45	Female	Control	20.00	12.53	Normal	Effect
3	43	Male	Drug B	17.33	10.93	Normal	Effect
4	59	Female	Control	22.67	14.67	Abnormal	No effect
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	54	Female	Drug B	16.80	11.73	Normal	Effect

1.1.2 Structure and feature of data

Any outcome of experiment or observation should be expressed with numerical data for statistical analysis. Most outcomes in medical research could be expressed through a data structure similar to Table 1.1, where 7 recorded items of 100 patients are given by a matrix with 100 rows and 7 columns. This is a basic format for data input in most of the statistical software such as SAS, SPSS, etc.

1.1.2.1 Basic observed unit

It is the basic unit for data collection determined by the purpose of research. For instance, if the systolic pressure and diastolic pressure are measured at a fixed time after treatment, then a patient is defined as an observed unit; otherwise, if the systolic pressure and diastolic pressure are measured at 3 different times after treatment (say, week 1, week 2 and week 4), then each patient is regarded as 3 observed units since the condition of each patient changes with time.

1.1.2.2 Recording item

The recording items used for statistical analysis usually consist of 3 parts: group, response variables and covariates. In Table 1.1, columns 2–8 show a 100×7 matrix corresponding to 7 recording items, of which treatment is a variable for grouping, systolic pressure, diastolic pressure, ECG and effectiveness are response variables, and age and gender are covariates.

1.2 Frequency Table and Histogram

Frequency table and histogram are not only fairly useful for description of sample data but also the intuitive foundation of the important concept of probability distribution.

1.2.1 Frequency table

As mentioned before, in a set of samples, the number of times a certain event occurs is frequency. For a complete list of mutually exclusive events, the table putting the corresponding frequencies together is called a frequency table.

1.2.1.1 Discrete-type frequency table

For a discrete variable, the completely and mutually exclusive events are just the possible values or categories of that variable. Based on the data of Example 1.2, two frequency tables are given in Tables 1.2 and 1.3,

Table 1.2 The frequency table for gender of 108 patients.

Gender	Frequency	Relative frequency (%)	Cumulative frequency	Cumulative relative frequency (%)
Female	45	41.7	45	41.7
Male	63	58.3	108	100.0
Total	108	100.0		

Table 1.3 The frequency table for occupation of 108 patients.

Occupation	Frequency	Relative frequency (%)	Cumulative frequency	Cumulative relative frequency (%)
Worker	28	25.9	28	25.9
Farmer	23	21.3	51	47.2
Businessman	24	22.2	75	69.4
Student	18	16.7	93	86.1
Soldier	15	13.9	108	100.0
Total	108	100.0		

where the ratio between the frequency and the total number is called relative frequency (if no confusion will arise, it is also called frequency). The sum of all relative frequencies must be 100% (in practice, sometimes it is not exactly 100% due to rounding error). The cumulative frequencies and cumulative relative frequencies are the results of successively cumulating the frequencies and relative frequencies respectively.

It is similar for ordinal variables. For instance, Table 1.4 is a frequency table for the results of certain semi-quantitative test among 150 patients; Table 1.5 is a frequency table for the treatment effect after their taking certain medicine.

1.2.1.2 Continuous type frequency table

For continuous variable, the general method to establish a frequency table could be learnt from the following example.

Table 1.4 The frequency table for the results of a semi-quantitative test among 150 patients.

Results	Frequency	Relative frequency (%)	Cumulative frequency	Cumulative relative frequency (%)
–	80	53.3	80	53.3
±	20	13.3	100	66.6
+	25	16.7	125	83.3
++	15	10.0	140	93.3
+++	10	6.7	150	100.0
Total	150	100.0		

Table 1.5 The frequency table for the treatment effect of certain medicine.

Effectiveness	Frequency	Relative frequency (%)	Cumulative frequency	Cumulative relative frequency (%)
Cure	65	43.3	65	43.3
Effect	45	30.0	110	73.3
No effect	25	16.7	135	90.0
Worse	15	10.0	150	100.0
Total	150	100.0		

Example 1.3 120 normal male adults were randomly selected from the residents of a county. Their red blood cell counts ($10^{12}/L$) were observed and listed as follows:

5.12	5.13	4.58	4.31	4.09	4.41	4.33	4.58	4.24	5.45	4.32	4.84
4.91	5.14	5.25	4.89	4.79	4.90	5.09	4.04	5.14	5.46	4.66	4.20
4.21	3.73	5.17	5.79	5.46	4.49	4.85	5.28	4.78	4.32	4.94	5.21
4.68	5.09	4.68	4.91	5.13	5.26	3.84	4.17	4.56	3.52	6.00	4.05
4.92	4.87	4.28	4.46	5.03	5.69	5.25	4.56	5.53	4.58	4.86	4.97
4.70	4.28	4.37	5.33	4.78	4.75	5.39	5.27	4.89	6.18	4.13	5.22
4.44	4.13	4.43	4.02	5.86	5.12	5.36	3.86	4.68	5.48	5.31	4.53
4.83	4.11	3.29	4.18	4.13	4.06	3.42	4.68	4.52	5.19	3.70	5.51
4.64	4.92	4.93	4.90	3.92	5.04	4.70	4.54	3.95	4.40	4.31	3.77
4.16	4.58	5.35	3.71	5.27	4.52	5.21	4.37	4.80	4.75	3.86	5.69

Try to establish a frequency table for this set of data.

Solution

- (1) Range R : The difference between the maximum and minimum of the data set is called the range. In our example, maximum = 6.18, minimum = 3.29, the range is $R = 6.18 - 3.29 = 2.89$.
- (2) Length of sub-intervals i : Divide the whole range into 8–15 sub-intervals. For convenience, take one tenth of the range first, and then slightly adjust to an easy number. In our example, $R/10 = 2.89/10 = 0.289 \approx 0.30$, then let $i = 0.30$.
- (3) Work out the list of sub-intervals: First of all, take a number slightly less than the minimum as the lower limit of the first sub-interval, say 3.20, such that its upper limit is $3.20 + 0.30 = 3.50$; take 3.50 as the lower limit of the second sub-interval such that its upper limit is $3.50 + 0.30 = 3.80$; Due to the fact that the upper limit of the former sub-interval is equal to the lower limit of the later one, for convenience, the upper limits are open and not shown except the last sub-interval, hence the list of sub-intervals are 3.20~, 3.50~, 3.80~, ..., 5.60 ~ and 5.90~6.20 (column 1 of Table 1.6).
- (4) Read, mark and count to get frequencies: Read over the data and write the five strokes of the Chinese character “正” one by one to mark and count the number of individuals corresponding to each sub-intervals (column 2 of Table 1.6).

Table 1.6 The frequency table based on the data set of red blood cell counts of 120 normal male adults.

Sub-interval	Mark	Frequency	Relative frequency (%)	Cumulative frequency	Cumulative relative frequency (%)
3.20–	丁	2	1.7	2	1.7
3.50–	正	5	4.2	7	5.9
3.80–	正正	10	8.3	17	14.2
4.10–	正正正正	19	15.8	36	30.0
4.40–	正正正正下	23	19.2	59	49.2
4.70–	正正正正正	24	20.0	83	69.2
5.00–	正正正正正	21	17.5	104	86.7
5.30–	正正正	11	9.2	115	95.9
5.60–	正	4	3.3	119	99.2
5.90–6.20	—	1	0.8	120	100.0
Total		120	100.0		

- (5) Calculate the frequencies, relative frequencies and cumulative frequencies (columns 3–6 of Table 1.6).

1.2.2 Frequency plot and histogram

To present the frequency table intuitively, a frequency plot within a coordinate system can be used, where the horizontal axis refers to “various situations” of the variable and the vertical axis refers to the corresponding frequencies.

1.2.2.1 Frequency plot for discrete variable — bar chart

For a discrete variable, one can use the points on the horizontal axis to express different categories or their related values; and plot vertical line segments on these points, of which the lengths express the frequencies or relative frequencies of the corresponding categories (Figs. 1.1 and 1.2). Such kind of frequency plot is called bar chart.

1.2.2.2 Frequency plot for continuous variable — histogram

For a continuous variable, one can use the sub-intervals with equal length on the horizontal axis to express the different situations of the variable; and

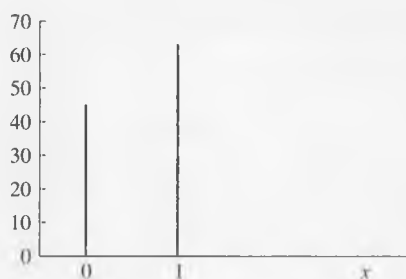


Fig. 1.1 The frequency plot for gender of 108 patients. x : gender, 0: female, 1: male.



Fig. 1.2 The frequency plot for occupation of 108 patients. y : occupation, 1: worker, 2: farmer, 3: businessman, 4: student, 5: soldier.

plot vertical rectangles on these intervals, of which the heights express the frequencies related to the sub-intervals (Fig. 1.3(a)), this is called histogram. However, when the lengths of the sub-intervals are not equal (for instance, the age intervals $0\sim$, $1\sim$, $5\sim$, $10\sim$, $15\sim$, ...), the heights cannot be used to express the frequencies.

Alternatively, one would use the areas of the rectangles to express the relative frequencies. The height of any rectangle in a histogram is neither the frequency nor the relative frequency, but the ratio of the relative frequency to the length of the sub-interval. Such kind of histogram is called frequency density histogram, of which the total area of all the rectangles is equal to 1 or 100%. The frequency density histogram can be used regardless of the lengths of the sub-intervals.

Both the frequency histogram and the frequency density histogram reflect the chances of various values taken by a continuous variable. The histograms in Fig. 1.3 appear to be symmetric, higher around the center

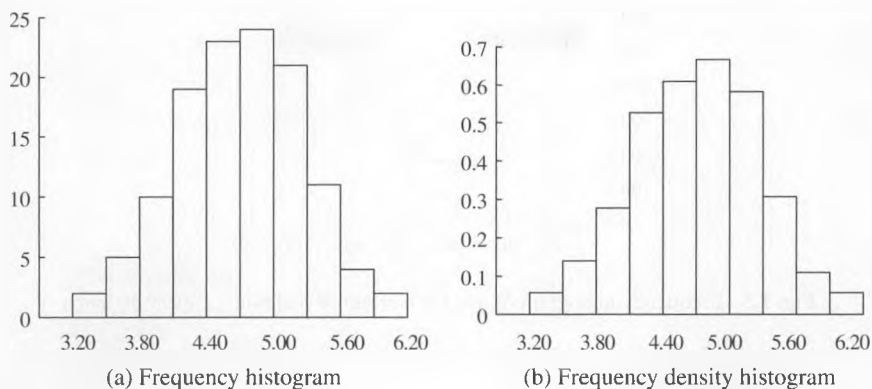


Fig. 1.3 Histograms plotted on the basis of the frequency table for the data set of red blood cell counts ($10^{12}/L$) of 120 normal male adults.

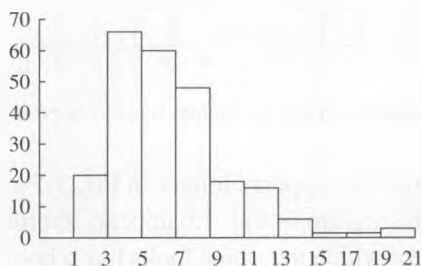


Fig. 1.4 Frequency histogram of hair mercury for the residents of a city.

and shorter on two sides, which indicate that the red blood cell counts of normal male adults, may be higher or lower with about equal chances, but mostly around the median level. Many histograms in practical problems look like this. However, there are some other types as well. For instance, the frequency histogram of hair mercury for the residents of a city is given in Fig. 1.4; the frequency histogram of the age for a group of male patients with lung cancer is given in Fig. 1.5; and the frequency histogram of the scores suggested by a group of patients for the importance of a specific item in evaluating the quality of life is given in Fig. 1.6. One can see that Figs. 1.4 and 1.5 are higher around center and shorter on two sides but not symmetric, of which the shape is usually called skew. The tail on the positive side appears longer in Fig. 1.4 and hence it is called positive skew; and

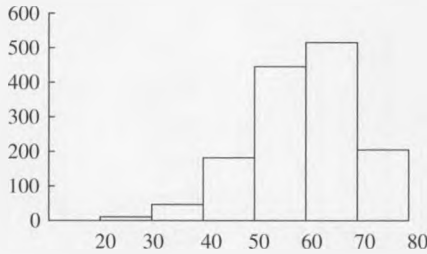


Fig. 1.5 Frequency histogram of age for a group of male lung cancer patients.

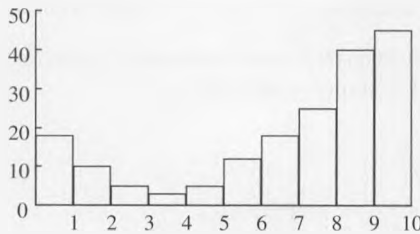


Fig. 1.6 Frequency distribution of the satisfactory score to an exhibition among the visitors.

the tail on the negative side appears longer in Fig. 1.5 and hence it is called negative skew. The histogram in Fig. 1.6 appears shorter around center and higher on two sides, of which the shape looks like a hook. Various shapes of the histograms are important for us to learn the distributions of continuous variables.

1.2.2.3 Frequency plot for ordinal variable — bar chart

The distances between successive ranks of an ordinal variable are usually not equal or unknown so that a bar chart instead of a histogram is used for frequency plot. For instance, the effect of a treatment can be described with four categories: cure, effect, no effect and worse, and the corresponding frequencies can be expressed with four bars on the horizontal axis as a bar chart for discrete variable.

1.2.3 Cumulative frequency plot

We can also use cumulative frequency plot to show how the frequency and percentage of individuals accumulate as the value increases, where

the horizontal axis refers to “various situations” of the variable and the vertical axis refers to the cumulative frequencies. According to Table 1.6, column 5 indicates the cumulative percentages at each observed red blood cell counts level among 120 normal male adults. We can get the cumulative frequency distribution based on the frequency table when using the data in column 5 as the values for vertical axes, and the upper limit values in column 1 as the values for horizontal axes (Fig. 1.7(a)). We can also get the cumulative frequency distribution based on the raw data (Fig. 1.7(b)). For example, there were 120 observations in Example 1.3, so each represents $1/120 = 0.83\%$. The first observation ($3.29 \times 10^{12}/L$) corresponds to a cumulative frequency of 0.83%, the first and second observations to a cumulative frequency of 1.67%, and so on. Cumulative frequency distribution is useful in finding the median and other quartiles. We can easily get the median, the lower and upper quartiles (25% and 75% quartiles) according to Fig. 1.7(b). The cumulative frequency distribution is a continuous ladder shape curve, say, the vertical jumps correspond to the increases in the cumulative frequencies at each observed red blood cell counts level. The cumulative frequency curve is steep when there is a concentration of values, and shallow when the values are sparse. In Fig. 1.7, the curve is steep in the center, and shallow around the low and high values. This means the majority of red blood cell counts are concentrated in the center of the distribution.

We usually use histograms to compare the distributions of two variables. However, cumulative frequency distribution provides more information when the histograms overlap extensively. For example, the histograms of the outcomes corresponding to the new medication group and control group mostly overlap in Fig. 1.8, thus we can hardly describe the difference between the two groups simply based on the histograms. In contrast, viewing at the cumulative frequency distribution, we can find that the change is sharper in the control group than that in the new medication group.

1.3 Measurement for Average Level of a Sample

In addition to frequency table and histogram, several numerical characteristics are also used for statistical description. For continuous variables, two often used characteristics are the average level and variation.

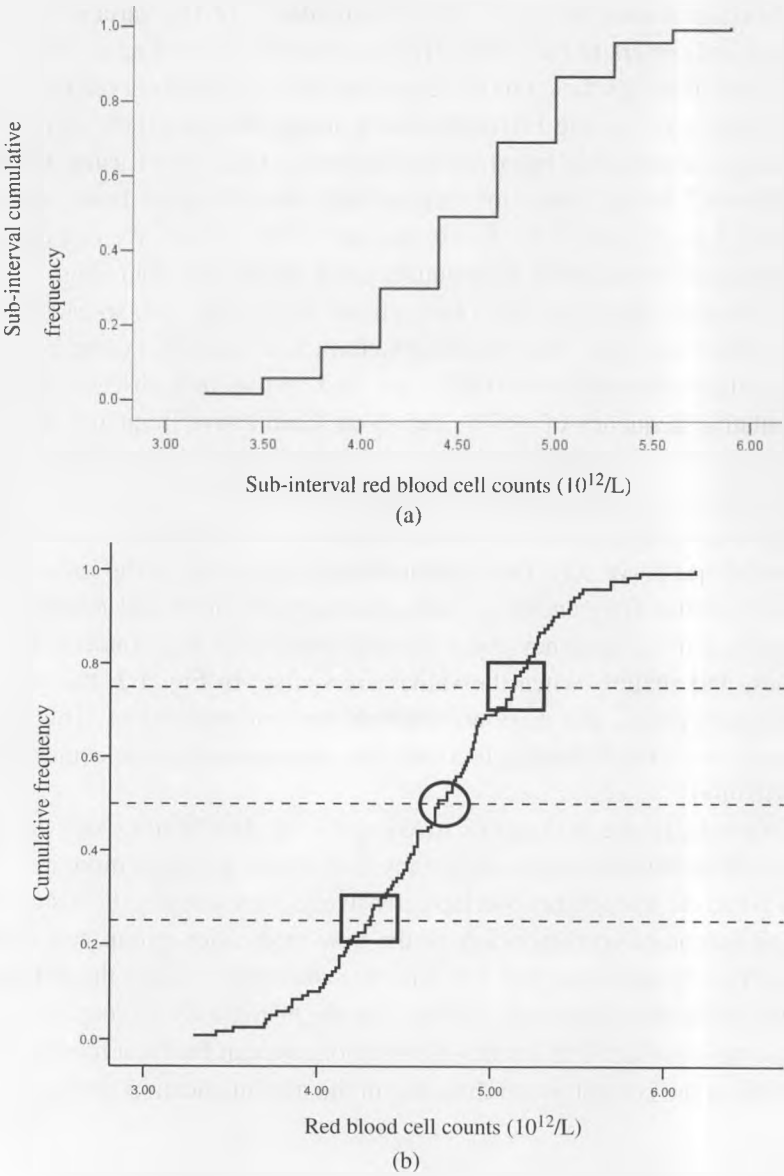


Fig. 1.7 Cumulative frequency distribution of the red blood cell counts ($10^{12}/L$) of 120 normal male adults.

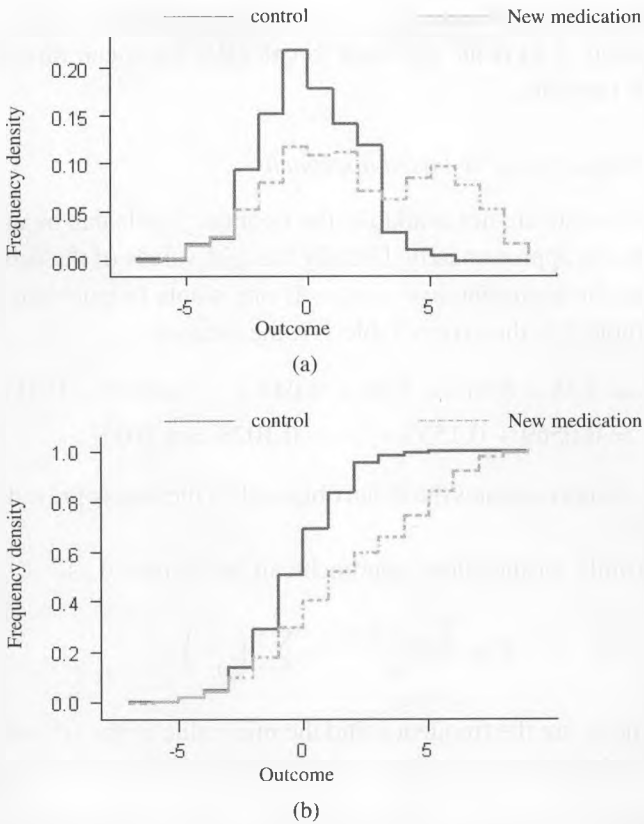


Fig. 1.8 Histograms and cumulative frequency distribution of effects outcome of the new medication group and the control group.

1.3.1 Arithmetic mean

When the histogram looks symmetric, the value that can well represent the average level is the arithmetic mean, or mean or average for brief, which is equal to the quotient of dividing the sum of observed values by the total number of individuals.

1.3.1.1 Raw data based approach

Denote the observed values of the individuals with x_1, x_2, \dots, x_n and the arithmetic mean with \bar{x} , then

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}, \quad (1.1)$$

Whenever no confusion arises, $\sum_{i=1}^n x_i$ could be simplified as $\sum_i x_i$ or even $\sum x$. Equation (1.1) is an approach to calculate the mean directly on the basis of the raw data.

1.3.1.2 Frequency table based approach

When the raw data are not available, the frequency table can be used to calculate the mean approximately. Usually the mid-values of the sub-intervals are taken as the representative values. If one wants to calculate the mean based on Table 1.6, then from Table 1.7, the mean is

$$\begin{aligned}\bar{x} &= 3.35 \times 0.017 + 3.65 \times 0.042 + \cdots + 6.05 \times 0.017 \\ &= 0.0569 + 0.1533 + \cdots + 0.1028 = 4.7057.\end{aligned}$$

Obviously, it approximates the mean obtained on the basis of raw data where $\bar{x} = 4.7167$.

The formula for the above approach can be expressed as

$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{n} = \sum_{i=1}^n \left(\frac{f_i}{n} \right) x_i, \quad (1.2)$$

where f_i and x_i are the frequency and the mid-value of the i th sub-interval, n is the total sample size. One can see from the process of the above

Table 1.7 The operation of weighted average based on a frequency table.

Sub-interval (1)	Mid-value (x) (2)	Frequency (f) (3)	Relative frequency (f/n) (4) = (3)/120	Mid-value \times Relative frequency (5) = (2) \times (4)
3.20–	3.35	2	0.017	0.0569
3.50–	3.65	5	0.042	0.1533
3.80–	3.95	10	0.083	0.3278
4.10–	4.25	19	0.158	0.6715
4.40–	4.55	23	0.192	0.8736
4.70–	4.85	24	0.200	0.9700
5.00–	5.15	21	0.175	0.9013
5.30–	5.45	11	0.092	0.5014
5.60–	5.75	4	0.033	0.1897
5.90–6.20	6.05	1	0.008	0.0484
Total		120	1	4.6939

calculation that the mid-value $x_6 = 4.85$ is multiplied by a bigger frequency $f_6/n = 20.0\%$ hence the contribution of x_6 is bigger. Such a way that the mid-values are not equally dealt with in the process of making average is called weighted average, and the result is called weighted mean. The relative frequency f_i/n in (1.2) that reflects the importance of the mid-value x_i is called weighting coefficient in general. The formula (1.2) is equivalent to the statement: the sample mean calculated based on a frequency table is a weighted mean of the mid-values with the frequencies as weighting coefficients.

1.3.2 Geometric mean

“Titer” is a widely applied measurement of concentration in microbiology and immunology where the tested material is proportionately diluted so that several samples with different concentrations are prepared and titrated respectively until certain phenomenon appears, of which the corresponding diluted proportion is defined as the measurement of the concentration. For instance, the concentrations of certain antibody are measured for a set of sample and the corresponding titers are 4, 8, 16, 16, 64, and 128, of which the arithmetic mean 39.3 is not an ideal representative of the data but the geometric mean. The arithmetic mean of the logarithms of the titers is calculated firstly,

$$(\log 4 + \log 8 + \log 16 + \log 16 + \log 64 + \log 128)/6 = 1.3045$$

then the anti-logarithm of it, $\log^{-1} 1.3045 = 20.16$, is the geometric mean of the above data set.

In general, if the individual values of the sample are all greater than 0, denoted with x_1, x_2, \dots, x_n , and the geometric mean is denoted with \bar{x}_g , then

$$\bar{x}_g = \log^{-1} \left(\frac{\log x_1 + \log x_2 + \dots + \log x_n}{n} \right) \quad (1.3)$$

or

$$\bar{x}_g = \sqrt[n]{x_1 x_2 \cdots x_n}. \quad (1.4)$$

When the histogram of the sample is positive skew, if the histogram of the logarithms is close to symmetric, then the geometric mean may well represent the average level and it is usually less than the arithmetic mean.

1.3.3 Median

When the histogram of the sample is taller around center and shorter on two sides but worse in symmetry, no matter positive skew or negative skew, the median, denoted with M_d , can be applied to measure the average level.

1.3.3.1 Raw data based approach

Arrange the individual values in the sample from smallest to largest; when the number of individuals n is an odd number, the observed value with rank $(n + 1)/2$ is taken as the median; when n is an even number, the average of the observed values with rank $n/2$ and $(n/2)+1$ is taken as the median. For example, the median of the data set $\{1, 1, 2, 2, 3, 4, 6, 9, 10\}$ is 3, while that of $\{1, 1, 2, 2, 3, 4, 6, 9, 10, 13\}$ is $(3 + 4)/2 = 3.5$.

1.3.3.2 Frequency table based approach

When only the frequency table is available, the median can be calculated approximately according to the following steps:

- (1) Calculate the rank corresponding to the median with $n/2$ approximately (may not necessarily be an integer);
- (2) Find out the sub-interval corresponding to the rank based on the cumulated frequencies, and denote with " $a \sim b$ " of which the length is $b - a$;
- (3) Find the cumulative frequencies up to the two ends of the sub-interval,
 f_a = the cumulative frequency of the last sub-interval
 f_b = the cumulative frequency of the current sub-interval
- (4) Estimate the value corresponding to the rank $n/2$ through interpolation

$$M_d \approx a + \frac{b - a}{f_b - f_a}(0.5n - f_a) \quad (1.5)$$

Example 1.4 The two columns of Table 1.8 is the frequency table related to Fig. 1.4. Calculate the arithmetic mean \bar{x} , geometric mean \bar{x}_g and median

Table 1.8 The frequency table of hair mercury (μ mol/kg) for the residents of a city.

Sub-interval	Frequency	Cumulative frequency	Mid-value (x)
1–	20	20	2
3–	66	86 (f_a)	4
5– ($a-b$)	60	146 (f_b)	6
7–	48	194	8
9–	18	212	10
11–	16	228	12
13–	6	234	14
15–	1	235	16
17–	1	236	18
19–21	3	239	20
Total	239		

M_d of hair mercury for the residents of the city approximately on the basis of these data.

Solution The 4th column of Table 1.8 is that of mid-values. The individual values are approximately equal to these mid-values respectively, and hence

$$\begin{aligned}\bar{x} &\approx (20 \times 2 + 66 \times 4 + 60 \times 6 + 48 \times 8 + \cdots + 3 \times 20)/239 \\ &= 1598/239 = 6.69 (\mu \text{ mol/kg})\end{aligned}$$

$$\begin{aligned}\bar{x}_g &\approx \log^{-1}(20 \times \log 2 + 66 \times \log 4 + 60 \times \log 6 + 48 \times \log 8 + \cdots \\ &\quad + 3 \times \log 20)/239 \\ &= \log^{-1}(0.7711) = 5.90 (\mu \text{ mol/kg})\end{aligned}$$

As for median, the corresponding rank is about

$$n/2 = 239/2 = 119.5$$

which is located in the sub-interval “5–7”; the cumulated frequency up to “5” (the cumulated frequency of the sub-interval “3–5”) is 86; the cumulated frequency up to “7” (the cumulated frequency of the sub-interval “5–7”) is 146; through interpolation,

$$M_d \approx 5 + \frac{7-5}{146-86}(119.5-86) = 6.12 (\mu \text{ mol/kg}).$$

1.4 Measurement for Variation of a Sample

In addition to the measure for average level, the measure for variation among individual values is also necessary. The four measures frequently used are introduced as follows.

1.4.1 Range R

It has been mentioned before that range is defined as the difference between the maximal value and the minimal value in the sample. Obviously, a bigger range indicates that the individual values are wider dispersed or higher varied. However, this measure depends on the maximal value and minimal value only but they often change a lot from sample to sample, and hence, R is worse in robustness.

1.4.2 $Q_3 - Q_1$

Arrange the n individual values in the sample from the smallest to largest; the value with a rank mostly close to $nP\%$ is called $P\%$ quartile or P percentile of the sample, denoted with X_p . As special cases, 50% quartile or 50th percentile is exactly the median; 25% quartile or 25th percentile is called the lower quartile, denoted with QL ; the 75% quartile or 75th percentile is called the upper quartile, denoted with QU .

The difference between QU and QL is another measure for variation. A bigger $QU - QL$ indicates that the individual values are wider dispersed. Here the information on ranks of the data is partly used, hence the robustness of $QU - QL$ is better than that of range R .

The raw data based approach for P th percentile is similar to that for median. Arrange the individual values in the sample from the smallest to largest. If $nP\%$ is an integer, then the value with this integer as rank is taken as the P th percentile. Otherwise, there are two integers closing to $nP\%$ and hence the average of the two corresponding values is taken as the P th percentile.

The steps of frequency table based approach for P th percentile are also similar to those for median, only that $n/2$ should be changed with $nP\%$,

$$X_p \approx a + \frac{b - a}{f_b - f_a}(nP\% - f_a). \quad (1.6)$$

1.4.3 Variance and standard deviation

Both the range and $Q_3 - Q_1$ share the common shortcoming that the individual information cannot be used sufficiently and the inference on variation of the population can hardly be performed.

The difference between individual value and the population mean is called deviation from the mean. It could be positive or negative though its absolute value reflects the variation. The average of squared deviations throughout the population is called the population variance, denoted by σ^2 , of which the dimension is square of the variable's dimension. To make the dimension same as that of the variable, square root of the population variance is defined as the population standard deviation, denoted with σ .

When the population mean is unknown and only the sample data are available, the population mean in the definition of deviation is replaced by the sample mean. It can be proved that the sum of the squared deviations from the sample mean must be less than that of the squared deviations from the population mean. To amend such a shortcoming, the sum is divided by $(n - 1)$ instead of n , and hence the average sum of squared deviations is called the sample variance, denoted with S^2 ,

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}, \quad (1.7)$$

where $n - 1$ is called the degrees of freedom. In fact, since the restrain of

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

among the n terms in the numerator of (1.7), there are only $n - 1$ deviations which could be varied freely.

For convenience in calculation, (1.7) can be expressed as

$$S^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2}{n - 1}. \quad (1.8)$$

The readers can easily prove the equivalence between (1.7) and (1.8) with elementary algebra.

The square-root of the sample variance is called the sample standard deviation, briefly denoted with S or SD , of which the dimension is the same as the variable itself. A bigger value of S refers to a greater variation.

1.4.4 Coefficient of variation

Sometimes the variations of two variables with different dimensions need to be compared. Obviously, their standard deviations cannot be compared directly because their dimensions are different. Then the coefficient of variation (CV), a measure without dimension, is useful, which is defined as

$$CV = \frac{S}{\bar{x}}. \quad (1.9)$$

Taking the height and weight of normal young males as an example, assume the mean and standard deviation of the height are 170 cm and 6 cm and those of the weight are 60 kg and 7 kg; their standard deviations 6 cm and 7 kg are not comparable while the comparison between their coefficients of variation $6/170 = 0.035$ and $7/60 = 0.117$ shows that the variation of weight is greater than that of height.

Mean and standard deviation are two important numerical characters for describing continuous variables so that conventionally they are often expressed together as $\bar{x} \pm s$. For instance, the above-mentioned mean and standard deviation of the variable of height could be expressed as 170 ± 6 (cm), where the symbol “ \pm ” just means “and”.

1.5 Relative Measures and Standardization Approaches

1.5.1 Ratio, frequency and intensity

In vital statistics and epidemiology, relative measures are widely used to describe the probability and intensity of certain event happening to the individuals in the population and often named with “... rate”. However, with careful consideration one will find that there are in fact three types of relative measures.

1.5.1.1 Ratio

It is simply a ratio of any quantity to another, such as

$$\text{Gender ratio of newly born babies} = \frac{\text{number of newly born girls}}{\text{number of newly born boys}}$$

and

$$\text{Mass index} = \frac{\text{Weight}}{\text{Height}^2} (\text{kg/m}^2),$$

where the numerator and denominator may not necessary be counted numbers nor of the same dimension.

1.5.1.2 *Relative frequency*

It is a special type of ratio where both the numerator and denominator are counted numbers and the numerator is part of the denominator. For a random sample, when the denominator is big enough, a relative frequency approximately describes the chance of certain event happening to the individuals in the population. For example, if 90 patients were cured among 100 treated ones, then

$$\text{Cure rate} = \frac{\text{number of cured}}{\text{number of treated}} = \frac{90}{100} = 90\%.$$

There is no dimension for relative frequency, and the value is a percentage or decimal within the interval of $[0,1]$.

1.5.1.3 *Intensity*

It is another special type of ratio where the denominator is the total observed person-years during certain period, the numerator is a number of certain event happening during the period. For example, the mortality rate is defined as

$$\begin{aligned} &\text{Mortality rate of certain year} \\ &= \frac{\text{number of deaths during the year}}{\text{person-years exposure to the risk of death during the year}}. \end{aligned}$$

The dimension of numerator is "person", that of denominator is "person \times year" so that the dimension of mortality rate is "person/(person \times year)" or "1/year". If the denominator is regarded as the "adjusted total number of persons \times 1 year", then the mortality rate can be regarded as the adjusted relative frequency per year.

In general, intensity as a type of relative measures could be understood as “relative frequency per unit of time”, reflecting the chance of certain event happening in a unit of time.

If an inference for a relative measure from sample to population is needed, one has to recognize the type of it, whether it is simply a ratio or a relative frequency or an intensity, because different type requires different statistical method.

1.5.2 Crude death rate and standardization

We will use the mortality rate as an example to show why the crude intensities are not directly comparable and how the standardization approaches work.

Table 1.9 gives two sets of data for two cities respectively, each of which includes several age groups; for each age group, the mid-year population, number of deaths during the year and age specific mortality rate are available. Ignoring the age groups and dividing the total number of deaths by the sum of mid-year populations, the crude mortality rates can be calculated, $P_a = 11.1\%$, $P_b = 23.3\%$. It seems that the risk of death in city *B* is higher than that in city *A*. However, in view of the age specific mortality rates, the risk of death in city *A* is higher than that in city *B* for all age groups. How to explain such a fallacy? Obviously, the crude mortality rate is incomparable because the distributions of age are not balanced between the two cities; it

Table 1.9 The data of age specific mortality rates for two cities.

Age group (year)	City A			City B		
	Mid-year population (10 ³)	Number of deaths (10 ³)	Mortality rate (%)	Mid-year population (10 ³)	Number of deaths (10 ³)	Mortality rate (%)
0~	400	2	5.0	288	1	3.5
15~	2000	10	5.0	238	1	4.2
30~	2000	15	7.5	794	5	6.3
45~	800	8	10.0	2000	18	9.0
60~	400	16	40.0	2000	70	35.0
75+	80	12	150.0	300	36	120.0
Total	5680	63	11.1	5620	131	23.3

is reasonable to compare the mortality rates age group by age group, but the variety of results based on separate comparisons can hardly be summarized into one conclusion.

A comprehensive measure summarizing the comparison between two sets of age specific mortality rates is often expected in applications such as comparison between different cities. There exist several methods for summary sharing a similar idea — standardization, that is, to adjust the imbalance in age distributions by selecting certain “standard” and calculating standardized mortality rates.

1.5.2.1 Direct standardization approach

The main steps of direct standardization are as follows: Select a “standard population” firstly; apply the whole set of age specific mortality rates to such a “standard population” and calculate the “expected number of deaths” for each age group in the “standard population”; calculate the crude mortality rate of the “standard population” based on the total expected numbers of deaths and call it a direct standardized mortality rate.

Example 1.5 Taking the sum of populations of the two cities in Table 1.9 as a “standard population”, compare the risk of death between the two cities through the direct standardization approach.

Solution Column 2 of Table 1.10 refers to the standard population which is the sum of the two populations for each age group; columns 3 and 5 refer

Table 1.10 Direct approach for standardized mortality rates of two cities.

Age group (year)	Standard population (10^3)	City A		City B	
		Mortality rate (%)	Expected number of deaths (10^3)	Mortality rate (%)	Expected number of deaths (10^3)
(1)	(2)	(3)	(4) = (2) × (3)	(5)	(6) = (2) × (5)
0–	686	5.0	3.43	3.5	2.40
15–	2238	5.0	11.19	4.2	9.40
30–	2794	7.5	20.96	6.3	17.60
45–	2800	10.0	28.00	9.0	25.20
60–	2400	40.0	96.00	35.0	84.00
75+	380	150.0	57.00	120.0	45.60
Total	11298	19.2	216.58	16.3	184.20

to the age specific mortality rates of the two cities respectively; columns 4 and 6 refer to the expected number of deaths for each age group if the mortality rate were applied to the "standard population" correspondingly; dividing the total expected numbers of deaths by the "standard population", one can obtain the direct standardized mortality rates for the two cities and put in the bottom cells of columns 3 and 5 respectively; and it concludes that the standardized mortality rate of city *A* is higher than that of city *B*. This is consistent with the conclusion obtained by age group comparison.

1.5.2.2 *Indirect standardization approach*

The main steps of indirect standardization are as follows: Select a set of "age specific mortality rates" as the "standard" first, apply it to the studied population and calculate the "expected number of deaths" for each age group of it; calculate the ratio between the total observed number of deaths and the total expected numbers of deaths and call it standard mortality ratio (SMR); multiplying the crude mortality rate of the "standard" with SMR, one can obtain the indirect standardized mortality rate for the studied population.

Example 1.6 Taking a set of age specific mortality rates as standard (see column 2 of Table 1.11), compare the risk of death between cities *A* and *B* based on the data in Table 1.9 through the indirect standardization approach.

Solution Columns 3 and 5 of Table 1.11 refer to the studied populations of the two cities; columns 4 and 6 refer to the expected numbers of deaths if the standard age specific mortality rates were applied to the studied populations respectively; dividing the total observed numbers of deaths (see columns 3 and 6 in Table 1.9) by the total expected numbers of deaths (see Table 1.11), one can obtain the SMRs for the two cities; multiplying the crude mortality rate of the "standard" with SMRs, one can obtain the indirect standardized mortality rates for cities *A* and *B* respectively.

Table 1.11 The indirect approach for standardized mortality rates of two cities.

Age group (year)	Standard mortality rate (%)	City A		City B	
		Mid-year population of City A (10^3)	Expected number of deaths in A (10^3)	Mid-year population of City B (10^3)	Expected number of deaths in B (10^3)
(1)	(2)	(3)	(4)=(2)×(3)	(5)	(6)=(2)×(5)
0–	4.3	400	1.72	288	1.24
15–	4.6	2000	9.20	238	1.09
30–	6.9	2000	13.80	794	5.48
45–	9.5	800	7.60	2000	19.00
60–	37.5	400	15.00	2000	75.00
75+	135.0	80	10.80	300	40.50
Total	17.2	5680	58.12	5620	142.31

$$\text{City A: SMR} = 63/58.12 = 1.084$$

$$\text{Indirect standardized mortality rate} = 17.2 \times 1.084 = 18.64(\%)$$

$$\text{City B: SMR} = 131/142.31 = 0.921$$

$$\text{Indirect standardized mortality rate} = 17.2 \times 0.921 = 15.84(\%)$$

Comparing the SMRs or the indirect standardized mortality rates between the two cities, one can find that the risk of death in city A is much higher than that in the city B.

1.5.2.3 Nature of crude mortality rate and standardized mortality rate

The crude mortality rate is a weighted average of age specific mortality rates with the sub-populations of age groups as the weight coefficients. If there are higher age specific mortality rates in the age groups with more populations, then the crude mortality rate is higher. Table 1.9 shows that the structures of populations in the two cities are obviously different, that is, more youths in city A but more elderly in city B. Therefore, offering higher weights to the higher age specific mortality rates, the weighted average results in a higher crude mortality rate of city B than that of city A.

In order to solve the problem of unequal weights, the idea of weighted average is still used in the direct standardization approach, but where the

sub-populations of age groups in the “standard population” are taken as the weights. Sometimes, different standard populations selected might result in quite different direct standardization mortality rates.

Totally giving up the information on age specific mortality rates, the indirect standardization approach keeps that on the numbers of deaths only. In fact, it is to calculate a weighted average of the selected standard age specific mortality rates with the observed sub-populations as the weights first; then SMR and use it to magnify or dwindle on the weighted average. Similarly, different sets of standard age specific mortality rates selected might result in quite different indirect standardization mortality rates.

The selection of standard populations or standard mortality rates is fairly important. Usually populations or mortality rates of the world or the country or the province are considered as the standard. If it is intended to compare two cities only, then the pool of the two populations or the pooled estimation of the age specific mortality rates (sum of the numbers of deaths in the age group/the sum of the sub-populations) might be taken as the standard. In practice, it is desirable to select more than one standard to see whether the results are consistent or not. If it is consistent, then the conclusion might be reliable; otherwise, one should be careful.

1.6 Frequently Used Graphs in Statistics

The first step of analysis is often to summarize and display the data, which can help us to identify outliers and possible errors in the data. Statistical chart is the important tool to display the data, which is intuitively clear by using the point-line-plane. There are several frequently used graphs in statistics, such as bar chart, percent bar chart, pie chart, line chart, semi-logarithmic line chart, box plot, and stem-and-leaf plot.

1.6.1 Layout of graphs

We can use Fig. 1.9 as an example to illustrate the layout of statistical graphs. The whole area for a graph is the chart area; the area within the X -axis and Y -axis is the drawing area; points and lines represent the original data. A two-dimensional graph consists of a horizontal coordinate axis (X -axis) and a vertical coordinate axis (Y -axis). For a three-dimensional graph, there is a third coordinate axis named Z -axis. There are scales on the

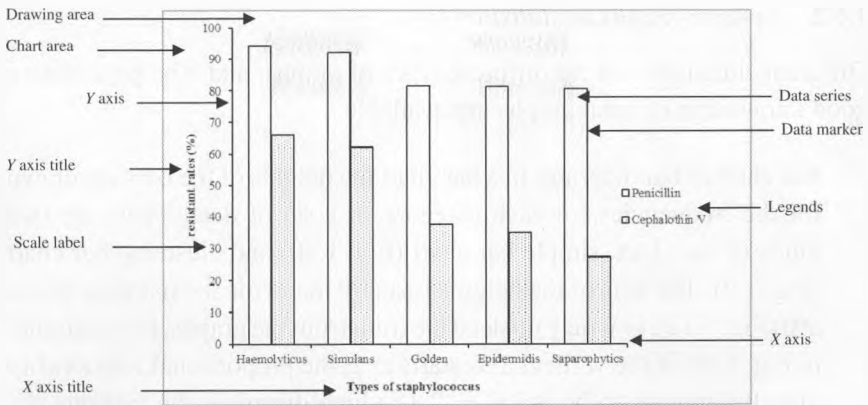


Fig. 1.9 Resistance rates of five types of staphylococcus for two kinds of antibiotics.

coordinate axes, and the corresponding numbers on the scales are named scale labels which could be real numbers or categories. Axes titles are left-aligned along the Y-axis and Z-axis or below the X-axis. There is no axes title for the pie chart. The basic rules for the layout of graphs are as follows:

- (1) Make sure that the graph is appropriate for the data and to support the main purpose of the study;
- (2) The title should be at the bottom of the graph;
- (3) Use different colors or different patterns for the different themes in the graphs, and put the legends at the appropriate place (at the right of the graph, bottom of the graph or top of the graph);
- (4) For the graphs which contain coordinate axes (bar chart, line chart, etc.), the values assigned for the X-axis should be in ascending order from the left to the right, and the values assigned for the Y-axis should be in ascending order from the bottom to the top. For the numerical variables, origin of coordinates, units and the appropriate scales should be labeled; for the categorical variables, the categories should be labeled. To make the graphs clearer, the height-width ratio is usually 5:7 (so-called "golden proportion").

There are several kinds of software to create the statistical graphs, such as Excel, SAS, SPSS, R, Maple, Matlab. We will list the SAS program for Fig. 1.9 in Sec. 1.7.

1.6.2 Several graphs in statistics

Different situations call for different types of graphs, and it helps to have a good knowledge of what graphs are available.

- (1) Bar chart or bar diagram: In a bar chart the heights of the bars are drawn for the frequencies for each category of a set of data. There are two kinds of bar chart, simple bar chart (Fig. 1.10) and clustered bar chart (Fig. 1.9). The axis of the heights (usually the vertical axis) must begin at 0 (Fig. 1.11), or it may mislead the truth from the graphs. For example, in Fig. 1.10, if the vertical axis starts at 2, the proportional relationship visually appears to be $A : B = 2 : 1$, which disguises the fact that the proportional relationship of A and B is 4:3. Each bar is in descending order of the variable in order to compare, and the space between two bars needs to be appropriate with a clear appearance.
- (2) Percent bar chart: The percent bar chart is used to display the frequency distribution. For example, Fig. 1.12 is plotted with the data in Table 1.12, where two bars with length equal to 100% are drawn for the two categories (hospitalization ≤ 7 days and > 7 days) at first

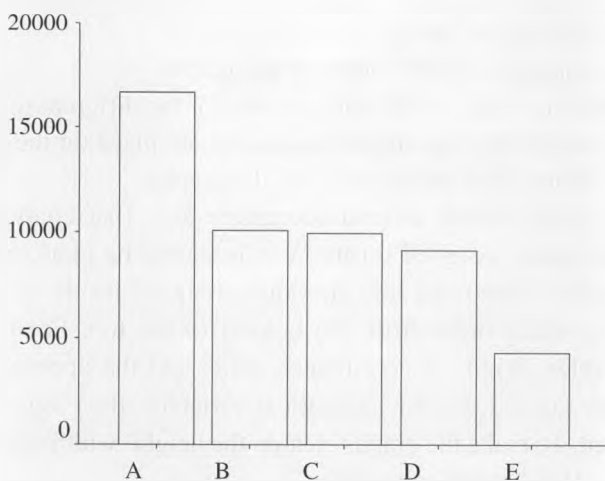


Fig. 1.10 Outpatient amount of the department of general internal medicine in the affiliated hospital of one medical university.

*A = Digestive; B = Cardiovascular; C = Respiratory; D = Endocrinology; E = Hematology.

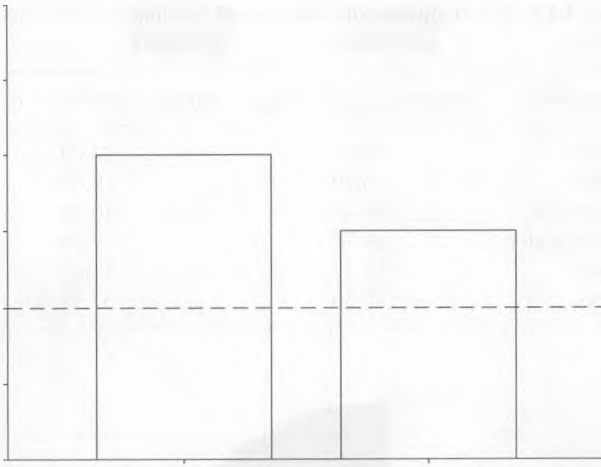


Fig. 1.11 The vertical axis must start at 0 in the bar chart.

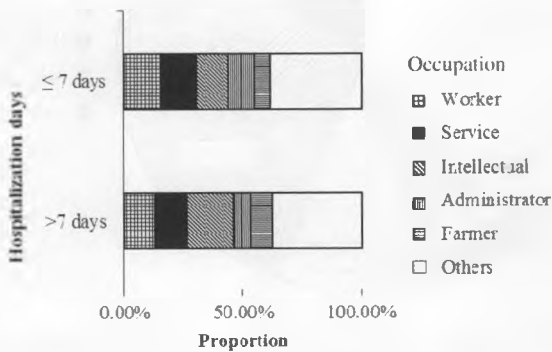


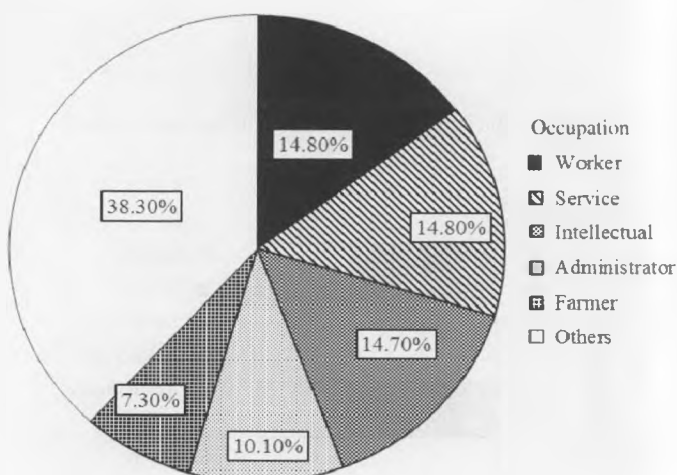
Fig. 1.12 The frequency distribution of the occupation of maternity patients with different hospitalization days.

step; and they are divided into several parts according to the proportion of the percentages of their component respectively in the second step. The sub-divided parts are sorted according to professional knowledge. If there is no exact sorting orders based on the professional knowledge, then they are usually in descending order by the proportion. The “other” is usually placed at the end of the bar.

- (3) Pie chart: The situation and the sorting orders for the pie chart are the same as the percent bar chart. The whole area (and consequently its

Table 1.12 The frequency distributions of occupations of maternity inpatients.

Occupation	Hospitalization ≤ 7 days	Hospitalization > 7 days
Worker	15.31	13.09
Farmer	6.79	9.06
Intellectual	13.40	19.46
Administrator	10.78	7.38
Service	15.22	13.42
Others	38.50	37.59

**Fig. 1.13** The frequency distribution of the occupation of maternity patients.

central angle) of the pie equal to 100%, and the circle is divided into sectors to illustrate the according proportions. The proportion of the first sector starts at the 12 o'clock position. If there is no professional concern, it is usually sorted from large to small value. Figure 1.13 uses the data of Table 1.13. It indicates the occupation distribution among 1402 maternity patients. It is obvious that the percent bar chart is better than pie chart when comparing proportions between multiple sets of data.

- (4) Line chart: The lines up and down in the rectangular plane coordinate system are used to display trends over time, or the changing process

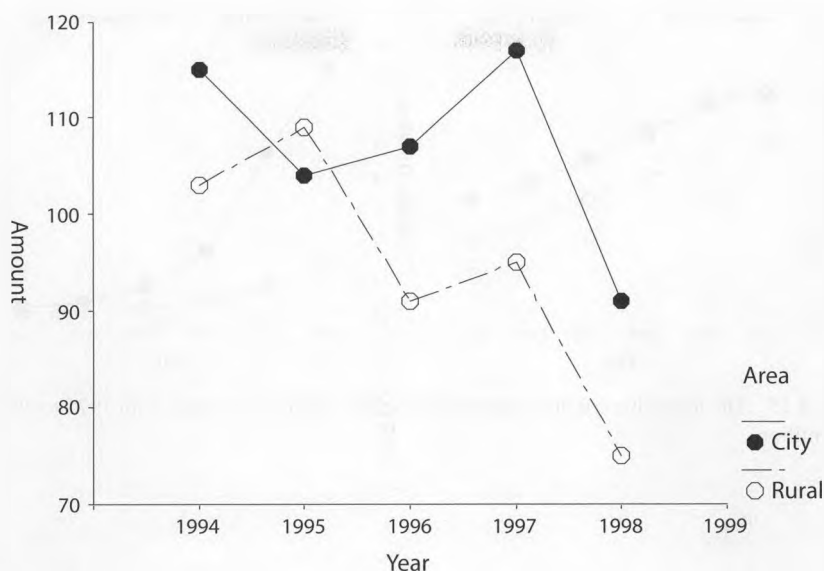


Fig. 1.14 The amount of discharged patients of the department of stomatology in the affiliated hospital of one medical university from 1994 to 1998.

Table 1.13 The mortalities of diarrhea and whooping cough (1/million) (1975–2000).

Year	Mortality	Diarrhea		Mortality	Whooping cough	
		Absolute decreased value (%)	Relative decreased value (%)		Absolute decreased value (%)	Relative decreased value (%)
1975	14.5			2.8		
1980	9.5	5.0	34.5	1.6	1.2	42.9
1985	3.7	5.8	61.1	0.9	0.7	43.8
1990	1.6	2.1	56.8	0.4	0.5	55.6
1995	0.7	0.9	56.3	0.2	0.2	50.0
2000	0.4	0.3	42.9	0.1	0.1	50.0

subject to the other things sequentially changing. The vertical and horizontal axes use the linear scale in the general line chart.

- (5) Box plot: It is useful to compare the average level and variation among different groups. It displays the minimum value, lower quartile (QL),

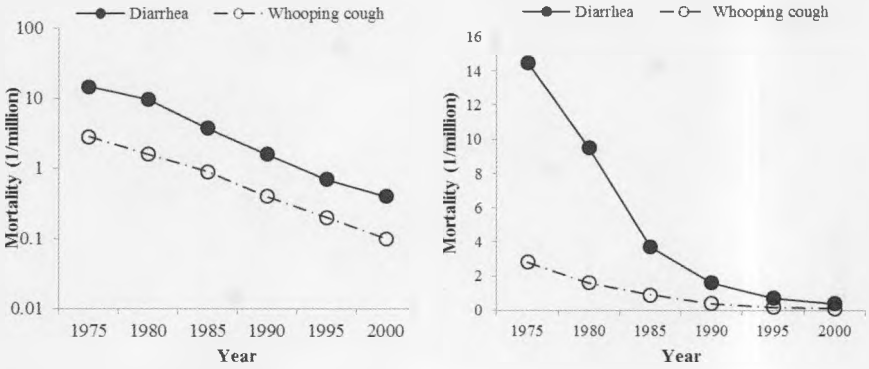


Fig. 1.15 The mortalities of diarrhea and whooping cough at one place from 1975 to 2000 (1/million).

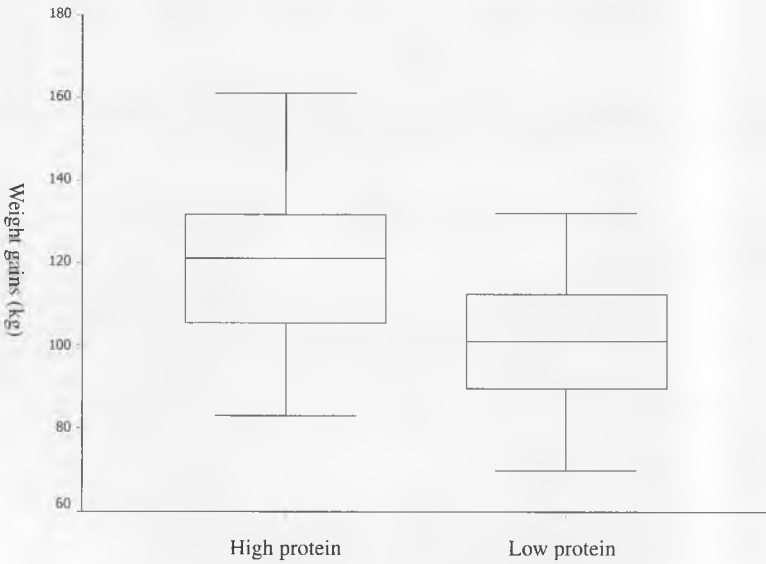


Fig. 1.16 The weight gains after using two kinds of feed with different protein content.

median (M), upper quartile (QU) and maximum value within each group. M indicates the average level; the range between minimum and maximum values, and the interquartile between QL and QU indicate the variation of the data.

- (2) Design the subgroups: By calculating the range and deciding the number of subgroups, it is obtained as follows:

Subgroup	Mid-value	Subgroup	Mid-value
3.20–	3.35	4.70–	4.85
3.50–	3.65	5.00–	5.15
3.80–	3.95	5.60–	5.75
4.10–	4.25	5.90–6.20	6.05
4.40–	4.55		

- (3) Organize data and list frequency table: Lines 08–21, each value is changed with the corresponding mid-value of its subgroup; lines 22–25 calculate description statistics such as mean, variance, standard deviation and variation coefficient (although the median and quartile could be given, they are just the mid-values in their sub-intervals instead

Program 1.1 Frequency table and histogram.

Line	Program	Line	Program
01	DATA RBC;	16	IF X < 4.70 & X >= 4.40 THEN
02	INPUT X @@;		Y=4.55;
03	CARDS;	17	IF X < 5.00 & X >= 4.70 THEN
04	5.12 5.13		Y=4.85;
05	18	IF X < 5.30 & X >= 5.00 THEN
06	3.86 5.69		Y=5.15;
07	;	19	IF X < 5.60 & X >= 5.30 THEN
08	PROC MEANS MIN MAX;		Y=5.45;
09	RUN;	20	IF X < 5.90 & X >= 5.60 THEN
10	DATA FRBC;		Y=5.75;
11	SET RBC;	21	IF X > 5.90 THEN Y=6.05;
12	IF X < 3.50 THEN Y=3.35;	22	PROC UNIVARIATE FREQ;
13	IF X < 3.80 & X >= 3.50 THEN	23	VAR Y;
	Y=3.65;	24	RUN;
14	IF X < 4.10 & X >= 3.80 THEN	25	PROC UNIVARIATE;
	Y=3.95;	26	CDF X / NORMAL; RUN;
15	IF X < 4.40 & X >= 4.10 THEN	27	PROC GCHART;
	Y=4.25;	28	VBAR Y/TYPE=PERCENT;
		29	VBAR Y/TYPE=CPERCENT;
		30	RUN;

of the values obtained by interpolation introduced above) and the frequency table is performed.

- (4) Histogram and cumulative frequency plot: Lines 25 and 26 work out the cumulative frequency plot, and lines 27–30 work out the frequency distribution and histogram.

Experiment 1.2 Clustered bar chart The detailed steps in the software for clustered bar chart are as follows (Program 1.2):

- (1) Data input: Lines 01–09, input the data as the following table: The resistance rates of five types of staphylococcus for two kinds of antibiotics

Types of staphylococcus	Penicillin (%)	Cephalothin V (%)
Golden	81.5	37.7
Epidermidis	81.3	35.1
Saprophytes	80.5	27.5
Haemolyticus	94.5	66.0
Simulans	92.3	62.1

- (2) Define the chart's structure: Lines 10–13 define the X-axis and Y-axis. X-axis is labeled as "Category of staphylococcus"; Y-axis is labeled as "RATE", ranged from 0% to 100%, 10% as the length of interval.
- (3) Label the variables: Lines 13–15, Label a1–a5 and a, b as the corresponding five categories of staphylococcus and two types of antibiotics, respectively.
- (4) Plot the chart: Line 16 is the plot PROC step; lines 17 and 18 indicate that the clustered bar chart should be plot based on the categories of staphylococcal bacteria and two types of antibiotics; line 19 defines the labels in step (3); lines 20–22 define the shape and the color in the chart.

Experiment 1.3 Pie chart The SAS codes for the pie chart is in Program 1.3. Firstly, line 01 sets the background of the chart, lines 02–12 input the data. Lines 13–15 make the frequency table of the dataset named JOB, and output the frequency table to the dataset named JOBPCT.

Program 1.2 Clustered bar chart.

Line	Program	Line	Program
01	DATA ANTIBIO;	13	PROC FORMAT;
02	INPUT BA\$ ANTIS RATE@@;	14	VALUE \$ss a1="G" a2="E"
03	CARDS;		a3="S" a4="H" a5="SM";
04	a1 a 81.5 a1 b 37.7	15	VALUE\$qq a="penicillin"
05	a2 a 81.3 a2 b 35.1		b="cephalothin V";
06	a3 a 80.5 a3 b 27.5	16	PROC GCHART;
07	a4 a 94.5 a4 b 66.0	17	WHERE ANTI in ("a", "b");
08	a5 a 92.3 a5 b 62.1	18	VBAR ANTI/GROUP=BA
09	;		SUMVAR=RATE
10	GOPTIONS RESET=ALL;		ATTENID=MIDPOINT;
11	AXIS1 LABEL=('staphylococcus')	19	FORMAT BA \$SS. ANTI \$QQ.;
	VALUE=('H SM G E S')	20	PATTERN1 V=L5 C=GRAY;
12	AXIS2 LABEL=('RATE')	21	PATTERN2 V=X5 C=GRAY;
	VALUE=(0 TO 100 BY 10);	22	RUN;

Program 1.3 Pie chart.

Line	Program	Line	Program
01	GOPTIONS RESET=ALL	19	PATTERN1 V=P3N0 C=GRAY;
	CBACK=WHITE BORDER	20	PATTERN2 V=E C=GRAY;
	HTITLE=12pt HTEXT 10pt;	21	PATTERN3 V=P3N45 C=GRAY;
02	DATA JOB;	22	PATTERN4 V=P3X45 C=GRAY;
03	LENGTH WORK \$8 ;	23	PATTERN5 V=P3N90 C=GRAY;
04	INPUT ID WORK;	24	PATTERN6 V=S C=GRAY;
05	DATALINES;	25	LEGEND1 LABEL=NONE
06	1 workers		POSITION=(RIGHT MIDDLE)
07	2 others		OFFSET=(,4) ACROSS=1
08	3 intellectual		VALUE=(COLOR=BLACK)
09	4 farmers		SHAPE=BAR(4,1.5);
10	5 services	26	PROC GCHART DATA=JOBPCT;
11	...	27	PIE WORK/SUMVAR=PERCENT
12	;		SLICE=INSIDE
13	PROC FREQ DATA=JOB;		PERCENT=INSIDE
14	TABLES WORK/OUT=JOBPCT;		LEGEND=LEGEND1
15	RUN;		MIDPOINTS='worker' 'others'
16	ODS RTF;		'farmers' 'managers' 'intellectual'
17	ODS GRAPHICS ON;		'services' NOHEADING;
18	TITLE 'JOB PERCENTAGE';	28	RUN;
		29	ODS GRAPHICS OFF;
		30	ODS RTF CLOSE;

Lines 16 and 17 define the output format of the pie chart as word file. Then line 8 defines the title, lines 19–24 define the patterns and the colors for each slice: V define THE dark or light and the patterns for each slice, and C defines the colors for each slice. Line 25 defines the legends: LEBEL defines the names of the legends, POSITION defines the positions of the legends, OFFSET defines the distance between the legends and the edge of the chart, ACROSS defines the amount of the legends (only one here), COLOR in VALUE defines the text color of the legends, and SHAPE defines the size of the legends.

Finally, lines 26–28 plot the pie chart based on the new dataset JOBPCT. MIDPOINTS define the slices are anti-clockwise ordered (PATTERN defines the slices' pattern according to the anti-clockwise order). Lines 29–30 complete the output.

Experiment 1.4 Box plot The SAS codes for the box plot is in Program 1.4. Line 01 sets the background for the plot; lines 02–07 input the data; lines 08–09 define the output format of the plot as word file, line 10 defines the title of the box plot. Line 11 defines the patterns of the plot: INTERPOL=BOXT5 defines the 95% percentile as the upper whisker and 5% percentile as the lower whisker. WIDTH defines the width of the box.

Program 1.4 Box plot.

Line	Program	Line	Program
01	GOPTIONS RESET=ALL CBACK=WHITE BORDER HTITLE=12pt HTEXT=10pt;	11	SYMBOL INTERPOL=BOXT5
02	DATA PROTEIN;	12	WITDTH=10; AXIS1 LABEL=NONE VALUE=(T=1 'high protein' T=2 'low protein')
03	INPUT GROUP \$ WEIGHT @@;		OFFSET= (5,5) LENGTH=50;
04	DATALINES;	13	AXIS2 LABEL= ('gain weight(g)') MINOR=NONE
05	A 134 A 146 A 104 A 119 A 124...		ORDER= (60 TO 180 BY20);
06	B 70 B 118 B 101 B 85 B 107...	14	PROC GLOT DATA=PROTEIN;
07	RUN;	15	PLOT WEIGHT*GROUP/HAXIS=AXIS1
08	ODS RTF;		VAXIS=AXIS2;
09	ODS GRAPHICS ON;	16	RUN;
10	TITLE1 'COMPARISON: WEIGHT BY GROUP';	17	ODS GRAPHICS OFF;
		18	ODS RTF CLOSE;

Program 1.5 Program for direct and indirect approaches.

Line	Program	Line	Program
01	DATA STA;	25	A2=P2*SP/1000;
02	INPUT P1 D1 P2 D2;	26	CARDS;
03	KEEP SP P1 R1 A1 A2;	27	4.3 400 286
04	R1=D1/P1*1000;	28	4.6 2000 238
05	R2=D2/P2*1000;	29	6.9 2000 794
06	SP=P1+P2;	30	9.5 800 2000
07	A1=R1*SP/11298;	31	37.5 400 2000
08	A2=R2*SP/11298;	32	135.0 80 300
09	CARDS;	33	;
10	400 2 286 1	34	PROC PRINT;
11	2000 10 238 1	35	PROC MEANS SUM
12	2000 15 794 5		NOPRINT;
13	800 8 2000 18	36	VAR A1 A2;
14	400 16 2000 70	37	OUTPUT OUT=STAN3
15	80 12 300 36		SUM=STA STB;
16	;	38	DATA STAN4;
17	PROC PRINT;	39	SET STAN3;
18	PROC MEANS SUM;	40	KEEP STA STB SMRA SMRB
19	VAR A1 A2;		SMPA SMPB;
20	RUN;	41	SMRA=63/STA;
21	DATA STA2;	42	SMRB=131/STB;
22	INPUT SP P1 P2;	43	SMPA=SMRA*17.2;
23	KEEP SP P1 P2 A1 A2;	44	SMPB=SMRB*17.2;
24	A1=P1*SP/1000;	45	PROC PRINT;
		46	RUN;

Lines 12 and 13 define the vertical and horizontal axis; lines 14–16 plot the box plot, finally lines 7–18 complete the output.

Experiment 1.5 Calculation of standardized mortality rate with direct and indirect approaches Program 1.2 is the SAS program for reference. The first 20 lines are for the direct approach where lines 4 and 5 calculate the age specific mortality rates, lines 7 and 8 calculate the age specific numbers of deaths, lines 10–17 list the data, and lines 18 and 19 calculate the standardized mortality rate.

Lines 21–46 are for the indirect approach where standardized mortality rates and sub-populations are required as input. Lines 24 and 25

calculate the age specific expected numbers of deaths; lines 27–34 list the data; lines 35–37 calculate the two total expected numbers of deaths respectively and put into STAN#; lines 41–44 calculate SMR and the standardized mortality rate respectively; then line 45 prints out the results.

1.8 Practice and Experiments

1. True or false: Which of the following statements are correct?
 - (1) “The red blood cells in occult blood examination” is a continuous variable.
 - (2) Red blood cell count is a discrete variable.
 - (3) The arithmetic mean is always greater than the median.
 - (4) The mean of large sample is always closer to the population mean than that of small sample.
 - (5) The arithmetic mean is always greater than the standard deviation.
 - (6) A histogram can be used to describe the distribution of the weight of a group of newborn babies.
 - (7) The cumulative frequency curve is a stepwise curve where the values are sparse.
 - (8) The distribution of the days of hospitalization for certain disease is higher around center and lower on two sides; the arithmetic mean is 10 days and the median is 5 days. One can see that the distribution is positive skew.
 - (9) The dimension of variation of coefficient is the same as that of the original variable.
 - (10) If the sample mean is greater, then the standard deviation must be greater.
 - (11) The range may increase with the increase of sample size.
2. Calculate the sample mean, median, variance, standard deviation and coefficient of variation for Example 1.4 on the basis of the raw data and the frequency table respectively; then compare and discuss the two sets of results.
3. The blood-glucose (mmol/L) is measured for 12 randomly selected patients. The data are 5.31, 6.12, 6.53, 6.53, 6.65, 6.66, 6.71,

- 6.93, 7.05, 7.15, 7.21, 7.35. Calculate the arithmetic mean, geometric mean and median; which answer better reflects the average level? Again calculate the range, $Q_3 - Q_1$ and standard deviation; which answer better reflects the variation?
4. The daily fat intake (g) of 100 randomly selected adults was surveyed with the data as follows:

23	60	78	84	90	104	114	127	130	143
43	69	81	94	97	102	117	120	147	150
52	80	88	96	103	105	114	128	130	153
65	79	89	95	107	108	128	131	139	148
67	75	76	91	102	105	127	138	153	167
70	72	95	103	111	117	128	130	147	142
67	62	72	95	109	111	127	132	144	151
23	37	69	88	99	109	119	139	134	155
30	89	76	96	93	104	117	133	147	151
44	73	83	94	96	107	111	128	131	150

Work out a frequency table, a histogram, box and whiskers plot, and stem and leaf plot; calculate the arithmetic mean, variance, standard deviation and coefficient of variation as well as median and $Q_3 - Q_1$.

5. Calculate the approximate arithmetic mean and standard deviation of red blood cell counts of 120 normal male adults based on the frequency table (Table 1.6) and compare with those calculated based on the raw data. Through this example, can you summarize the main steps for calculating arithmetic mean and standard deviation based on a frequency table in general?
6. It is quite popular to use two different concepts to describe the incidence of disease:

Cumulated incidence rate

$$= \frac{\text{Number of new patients during the same period}}{\text{Total number of persons followed during certain period}}$$

Person-year incidence rate

$$= \frac{\text{Number of new patients during the same period}}{\text{Total person-years of exposure to the risk during certain period}}$$

Discuss the properties of these two rates; are they ratio, frequency or intensity?

7. The data of liver-cancer specific mortality rates for males in two cities are collected as follows (Gong Zhiping, 1992):

Age group (year)	City A				City B			
	Population	Proportion	Number of deaths	Mortality rate	Population	Proportion	Number of deaths	Mortality rate
0~	323600	0.6555	24	7.4	364500	0.6949	22	6.0
30~	56800	0.1150	75	132.0	64300	0.1226	75	116.6
40~	42400	0.0859	103	242.9	40100	0.0765	104	259.4
50~	30500	0.0618	87	285.2	28800	0.0549	84	291.7
60~	21300	0.0431	69	323.9	16200	0.0309	54	333.3
70~	19100	0.0387	33	172.8	10600	0.0202	22	207.5
Total	493700	1.0000	391	79.2	524500	1.0000	361	68.8

Compare the risk of liver cancer between the two cities through the direct standardization approach.

- (1) Taking the population of city A as a standard population;
 - (2) Taking the population of city B as a standard population;
 - (3) Taking the total population of cities A and B as a standard population;
 - (4) Compare and discuss the results.
8. Compare the risk of liver cancer between the two cities through the indirect standardization approach.
- (1) Taking the age specific mortality rates of city A as standard mortality rates;
 - (2) Taking the age specific mortality rates of city B as standard mortality rates;
 - (3) Taking the pooled age specific mortality rates of cities A and B as a standard population;
 - (4) Compare and discuss the results.
9. What are the frequently used graphs in statistics? What are the different situations for the use of different types of graphs?

10. Prove or check the following statements. Assume there are observed values y_1, y_2, \dots, y_n , and denote $\bar{y} = \sum_{i=1}^n \frac{y_i}{n}$.

$$(1) \sum_{i=1}^n ay_i = a \sum_{i=1}^n y_i; \quad (2) \sum_{i=1}^n (y_i - \bar{y}) = 0;$$

$$(3) \sum_{i=1}^n (a + y_i) = na + \sum_{i=1}^n y_i; \quad (4) \sum_{i=1}^n \left(\frac{y_i}{n} a \right) = a \sum_{i=1}^n \frac{y_i}{n};$$

$$(5) \sum_{i=1}^n (y_i + a)^2 = \sum_{i=1}^n y_i^2 + 2a \sum_{i=1}^n y_i + na^2;$$

$$(6) \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2;$$

$$(7) \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2.$$

(1st edn. Jiqian Fang; 2nd edn. Chun Hao, Jiqian Fang)

Chapter 2

Probability and Distribution

One of the main tasks of statistics is inference from sample to population. As it is inference, there must be risk so that one has to clarify to what degree of accuracy their conclusion of inference can reach, but not to use ambiguous language such as “probably”. The regular and scientific way of statement in statistics is to give the probability P incorporating in the conclusion. How to determine the value of P ? The probability distributions of various types of variables should be studied, which themselves not only have important medical applications, but also play the role of theoretical foundation for the successive chapters in this book.

2.1 Explanation of Probability and Related Concepts

2.1.1 *Probability*

After an ideally uniform coin is flipped, either the side with head or without may appear equal-likely so that, we may say, the probability of the event “head up” is equal to 0.5; after an ideally uniform die with figures of 1, 2, ..., 6 is tossed, any one of the figures may appear equal-likely so that the probability of the event “3 up” is equal to 1/6. It sounds easy but in practice it is not so smooth.

Before a color blindness test, any student might be and might not be color blind, not equal-likely or easy to get the probabilities. Usually, after a randomly selected group of n students is tested, if m of them are color blind, then the frequency m/n can be regarded as the approximation of the probability of color blindness in the population of students.

Besides the model-based probabilities in the problems such as coin flipping and die tossing and the frequency-based probabilities in the problems such as color blindness, there is alternative kind of so-called

subjective probability. For instance, when a patient feels chest distress and pain, the physician may think that the probability of the event "heart disease" occurring is 20%. As a matter of fact, the situation, whether with or without heart disease, has already been determined so that such a "20%" is just to describe the belief of the physician on the event of "heart disease".

One can see, for the same word "probability" there are various understandings, which may sometimes lead to confusion. In this volume, we are only concerned with frequency-based probabilities.

In general, the possible outcomes of a trial, denoted by E_1, E_2, \dots , are called events. The chance of certain event E occurring in a trial is called the probability of the event E , denoted by $P(E)$. Any probability takes value from the interval $[0, 1]$; the event with probability 0 is called impossible event; the event with probability 1 is called certain event; and the event with probability between 0 and 1 is called random event.

For any two events E_1 and E_2 , under the condition that E_1 occurs, the probability of the event E_2 occurring is called conditional probability, denoted as $P(E_2|E_1)$, where E_1 is the condition and E_2 is the event. For instance, the probability of "nasopharyngeal carcinoma" of a patient with EB virus positive is a conditional probability, denoted as

$$P(\text{with nasopharyngeal carcinoma} | \text{EB virus positive})$$

2.1.2 Odds

If any two events E_1 and E_2 are not possible to appear simultaneously, they are called exclusive events. As a kind of special cases, if there are only two possible events and they are exclusive, denoted with E and \bar{E} , then they are called complementary events, and obviously

$$P(\bar{E}) = 1 - P(E). \quad (2.1)$$

The ratio between the probability of any event E and that of its complementary event is called odds,

$$\text{Odds} = \frac{P(E)}{P(\bar{E})} = \frac{P(E)}{1 - P(E)}. \quad (2.2a)$$

An odds greater than 1 indicates that the event E is dominant compared to \bar{E} ; an odds equal to 1 indicates that the events E and \bar{E} are nip and tuck.

For example, if the incidence rates of influenza in classes A , B and C are 60%, 50% and 40%, then the odds are 1.50, 1.00 and 0.67, respectively. Conversely, the probability of E can be derived from its odds as well, based on (2.2a),

$$P(E) = \frac{\text{Odds}}{1 + \text{Odds}} \quad (2.2b)$$

2.1.3 Bayes' formula

The public is interested to know the relationship between lung cancer (B) and smoking (A). Ideally, to study such a topic, it is better to randomly divide the subjects into two groups, then invite one group to smoke and forbid the other group; follow up year by year to obtain the conditional probability of "lung cancer" under the condition of "smoking", $P(B|A)$. Unfortunately, it is infeasible. Usually, we randomly select a group of patients with lung cancer and ask about their history of smoking to get a conditional probability of "smoking" under the condition of "lung cancer", $P(A|B)$. How do we derive $P(B|A)$ (the conditional probability of "lung cancer" under the condition of "smoking") on the basis of $P(A|B)$ (the conditional probability of "smoking" under the condition of "lung cancer")? It is a significant problem in practice. The following is the key formula to solve such a problem (See Sec. 2.6 for the proof)

$$P(B|A) = \frac{P(B)P(A|B)}{P(B)P(A|B) + P(\bar{B})P(A|\bar{B})} \quad (2.3)$$

It is called Bayes' formula, where \bar{B} is the complementary event of B .

Example 2.1 The percentage of people suffering from lung cancer in a population is 20×10^{-5} . The percentage of "smokers smoke" (A) among the patients "with lung cancer" (B) is 80%, but that among normal people "without lung cancer" (\bar{B}) is 16%. What is the probability that a person suffers from lung cancer among smokers? What is the odds that a person suffers from lung cancer in the population? What is the odds that a person suffers

from lung cancer among smokers? What is the ratio between the latter and the former and what does the ratio really mean?

Solution $P(B) = 20 \times 10^{-5}$, $P(A|B) = 0.8$, $P(A|\bar{B}) = 0.16$. By (2.3), the probability that a person suffers from lung cancer among smokers is

$$\begin{aligned} P(B|A) &= \frac{(20 \times 10^{-5})(0.8)}{(20 \times 10^{-5})(0.8) + (1 - 20 \times 10^{-5})(0.16)} \\ &= 99.92 \times 10^{-5} \end{aligned}$$

The odds that a person suffers from lung cancer in the population is

$$\frac{P(B)}{P(\bar{B})} = \frac{20 \times 10^{-5}}{1 - 20 \times 10^{-5}} = 2.0004 \times 10^{-4}.$$

The odds that a person suffers from lung cancer among smokers is

$$\frac{P(B|A)}{P(\bar{B}|A)} = \frac{99.92 \times 10^{-5}}{1 - 99.92 \times 10^{-5}} = 10.002 \times 10^{-4}.$$

If there are two odds' for the same event under different conditions, one may use the ratio between the two odds' to compare the risks under the two conditions. In this example, the ratio between the latter and the former is

$$\text{Odds ratio} = \frac{P(B|A)}{P(\bar{B}|A)} \div \frac{P(B)}{P(\bar{B})} = \frac{10.002 \times 10^{-4}}{2.0004 \times 10^{-4}} = 5.00.$$

This reflects the risk of smoking leading to lung cancer.

The risk can also be described with another concept called likelihood ratio, which is defined as the ratio of two conditional probabilities of the same event under the conditions that two complementary events appear respectively.

$$\text{Likelihood ratio} = \frac{P(A|B)}{P(A|\bar{B})} = \frac{0.8}{0.16} = 5.00$$

This example shows that

$$\text{Odds ratio} = \text{Likelihood ratio}. \quad (2.4)$$

In fact, this is a generally held statement and a theoretical foundation of retrospective studies. The proof will be given in Sec. 2.6.

2.2 Distributional Characters of Random Variables

2.2.1 Probability function of discrete random variable

The outcome of flipping a uniform coin can be regarded as a random variable, which takes a “value” either “head up” or “head down” with probabilities

$$P(\text{head up}) = 0.50, \quad P(\text{head down}) = 0.50.$$

The outcome of tossing a uniform die can be regarded as a random variable, which takes one and only one value out of 1, 2, 3, 4, 5, and 6 with probabilities

$$P(1 \text{ up}) = P(2 \text{ up}) = \cdots = P(6 \text{ up}) = \frac{1}{6}.$$

In general, there are two aspects for a random variable: the possible values and the corresponding probabilities. To describe a discrete random variable, the possible values should be listed and the probability of each value should be given; these two as a whole are called probability function, denoted by $P(X)$.

Example 2.2 Taking balls out from a pocket There is a big pocket containing many small balls with the same size and weight, of which 80% are black and 20% are white. Take out one ball after stirring and mixing up, record its color; send it back to the pocket, stir and mix, take one out again ... In such a way, after repeating for five times, the total number of times of black ball appearing is a discrete random variable. It will be proved that the probability function can be shown as Table 2.1 and Fig. 2.1.

Table 2.1 The probability function of the discrete random variable in Example 2.2.

The possible values of X	0	1	2	3	4	5
Probability $P(X)$	0.0003	0.0064	0.0512	0.2048	0.4096	0.3277

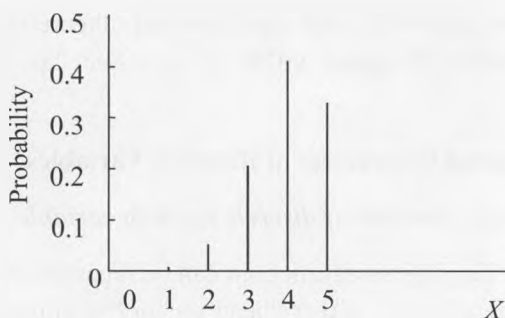


Fig. 2.1 The plot of the probability function corresponding to Table 2.1.

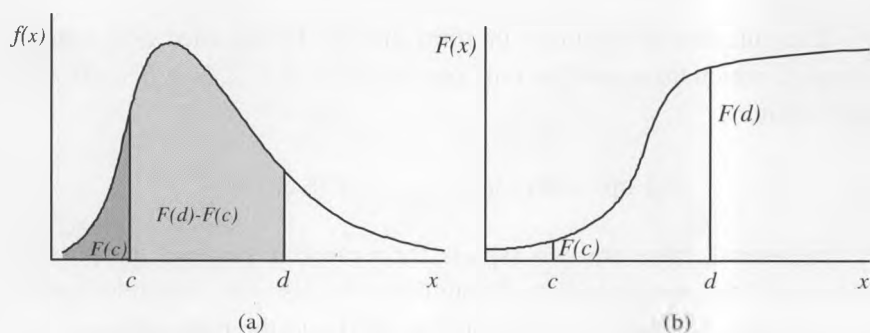


Fig. 2.2 Relationship between (a) probability density function and (b) probability distribution function.

2.2.2 Probability density function of continuous random variable

For continuous random variables, such as red blood cell counts and hair mercury content, the possible values can hardly be listed completely. When the sample size is considerably large, the frequency density plot can approximately reflect the distributional character; one may imagine, when the sample size increases and the length of the sub-intervals decreases infinitely, the profile of the frequency density plot will tend to be a smooth curve, which is called probability density curve (Fig. 2.2(a)). The related function of such a curve is called probability density function, denoted by $f(x)$.

2.2.3 Distribution function of random variable

Similar to cumulated relative frequency, there is a concept of cumulated probability for any random variable (including discrete one), which is equal to the probability of the event that the value of the random variable X is less than or equal to x , denoted by

$$F(x) = P(X \leq x).$$

This function will not decrease with the increase of x but reflects how the cumulated probability changes with x so that it is called cumulative probability function or distribution function. Obviously, any type of random variables, no matter discrete or continuous or ordinal ones, can share the concept of distribution function. This brings much convenience to the research and application of mathematical statistics.

Figure 2.2 demonstrates the relationship between probability density function and distribution function. Figure 2.2(a) is a plot of a distribution function, of which the shape may vary from problem to problem in practice but the area under the curve (above the x -axis) is always equal to 1; the area of the left tail up to c under the curve (above the x -axis) is just the cumulative probability $P(X \leq c)$, of which the value is equal to $F(c)$, the height of the curve at $x = c$ in Fig. 2.2(b). The difference between $F(d)$ and $F(c)$ in Fig. 2.2(b) is equal to the probability of the event that the variable X falls in the interval of (c, d) , that is, the area of the mid-part between c and d under the curve (above the x -axis) in Fig. 2.2(a).

2.2.4 Population mean and population variance

Similar to the concepts of sample mean and sample variance, the population mean and population variance describe the average level and variation of the population, respectively.

To any discrete random variable X , denoting the possible values and related probabilities with x_1, x_2, \dots, x_k and p_1, p_2, \dots, p_k respectively (k could be infinity as well), if (2.5) results in a finite number, then it is called population mean of X or expectation of X , denoted by $E(X)$,

$$E(X) = \sum_{i=1}^k p_i x_i = \mu. \quad (2.5)$$

Similarly, if (2.6) results in a finite number, then it is called population variance of X , denoted with $Var(X)$,

$$Var(X) = \sum_{i=1}^k p_i (x_i - \mu)^2 = \sigma^2. \quad (2.6)$$

In fact, $E(X)$ is tantamount to a weighted average of all possible values throughout the population; and $Var(X)$ is tantamount to a weighted average of all possible squared deviations from the mean μ throughout the population; the weighted coefficients used are the related probabilities.

To any continuous random variable X , denoting the domain with (a, b) and the probability density function with $f(x)$, if (2.7) results in a finite number, then it is called population mean of X or expectation of X , denoted by $E(X)$,

$$E(X) = \int_a^b x f(x) dx = \mu, \quad (2.7)$$

where dx is tantamount to the length of a sub-interval, $f(x)dx$ is tantamount to the probability of the event that the variable X falls in the interval of $(x, x + dx)$, the integral sign \int_a^b is tantamount to the sign of $\sum_{i=1}^k$ in (2.5). Similarly, if (2.8) results in a finite number, then it is called population variance of X , denoted with $Var(X)$,

$$Var(X) = \int_a^b (x - \mu)^2 f(x) dx = \sigma^2. \quad (2.8)$$

In fact, (2.7) is still tantamount to a weighted average of all possible values throughout the population, and (2.8) is still tantamount to a weighted average of all possible squared deviations from the mean μ throughout the population.

For both discrete and continuous variables, there are several important properties of $E(X)$ and $Var(X)$, which may be useful later.

(1) If a is a constant, then

$$E(aX) = aE(X), \quad (2.9)$$

$$Var(aX) = a^2 Var(X). \quad (2.10)$$

(2) For any two random variables X_1 and X_2 ,

$$E(X_1 + X_2) = E(X_1) + E(X_2). \quad (2.11)$$

(3) When and only when X_1 and X_2 are independent of each other,

$$\text{Var}(X_1 \pm X_2) = \text{Var}(X_1) + \text{Var}(X_2). \quad (2.12)$$

(4) For any random variable X ,

$$\text{Var}(X) = E(X^2) - [E(X)]^2. \quad (2.13)$$

2.3 Binomial Distribution

2.3.1 Probability function

Let us return to Example 2.2. Table 2.1 gives the probability function of the random variable X , which is the total number of times black ball appears out of five times of repeated sampling with replacement of ball from the pocket. The values of this probability function are calculated in Table 2.2. Obviously, they are just corresponding to the terms of the following expansion.

$$\begin{aligned} (0.2 + 0.8)^5 &= \binom{5}{0} (0.2)^5 + \binom{5}{1} (0.8)(0.2)^4 + \binom{5}{2} (0.8)^2(0.2)^3 \\ &\quad + \binom{5}{3} (0.8)^3(0.2)^2 + \binom{5}{4} (0.8)^4(0.2) + \binom{5}{5} (0.8)^5. \end{aligned}$$

Here $\binom{n}{x}$ refers to the number of possible combinations after picking up x items from n ,

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}, \quad 0! = 1, \quad k! = k(k-1) \cdots (2)(1), \quad k \neq 0.$$

In general, if the probability of an event appearing in a trial is π , and after n times of independently repeated trials, the total number of times such an event appears is a random variable, denoted by X ; then the probability

Table 2.2 Derivation of the probability function in Example 2.2.

Total number of times black ball appears X	Possible outcome* (permutation)	Probability of permutation	Probability of combination $P(X)$
0	OOOOO	$(0.2)^5$	$\binom{5}{0} (0.2)^5 = 0.0003$
1	XOOOO OXOOO OOXOO OOOXO OOOOX	$(0.8)(0.2)^4$ $(0.8)(0.2)^4$ $(0.8)(0.2)^4$ $(0.8)(0.2)^4$ $(0.8)(0.2)^4$	$\binom{5}{1} (0.8)(0.2)^4 = 0.0064$
2	XXOOO XOXOO XOOXO XOOOX : OOOXX	$(0.8)^2(0.2)^3$ $(0.8)^2(0.2)^3$ $(0.8)^2(0.2)^3$ $(0.8)^2(0.2)^3$: $(0.8)^2(0.2)^3$	$\binom{5}{2} (0.8)^2(0.2)^3 = 0.0512$
3	XXXOO XXOXO XXOOX XOXXO : OOXXX	$(0.8)^3(0.2)^2$ $(0.8)^3(0.2)^2$ $(0.8)^3(0.2)^2$ $(0.8)^3(0.2)^2$: $(0.8)^3(0.2)^2$	$\binom{5}{3} (0.8)^3(0.2)^2 = 0.2048$
4	XXXXO XXXOX XXOXX XOXXX OXXXX	$(0.8)^4(0.2)$ $(0.8)^4(0.2)$ $(0.8)^4(0.2)$ $(0.8)^4(0.2)$ $(0.8)^4(0.2)$	$\binom{5}{4} (0.8)^4(0.2) = 0.4096$
5	XXXXXX	$(0.8)^5$	$\binom{5}{5} (0.8)^5 = 0.3277$
Total			1.0000

*O: white ball, X: black ball.

of $X = x$ can be calculated as follows:

$$P(x) = \binom{n}{x} (1 - \pi)^{n-x} \pi^x, \quad x = 0, 1, \dots, n. \quad (2.14)$$

Since (2.14) is the same as the $(x + 1)$ th term of Newton's binomial expansion

$$\begin{aligned} [(1 - \pi) + \pi]^n &= \binom{n}{0} (1 - \pi)^n + \binom{n}{1} (1 - \pi)^{n-1} \pi \\ &+ \binom{n}{2} (1 - \pi)^{n-2} \pi^2 + \dots + \binom{n}{x} (1 - \pi)^{n-x} \pi^x \\ &+ \dots + \binom{n}{n-1} (1 - \pi) \pi^{n-1} + \binom{n}{n} \pi^n, \end{aligned}$$

it is called the probability function of binomial distribution, and the random variable X is called a binomial variable, or a variable following a binomial distribution, denoted by

$$X \sim B(\pi, n).$$

Obviously, the distribution function of binomial variable is in the shape of

$$P(X \leq x) = P(0) + P(1) + \dots + P(x) = \sum_{k=0}^x P(k). \quad (2.15)$$

Example 2.3 The 50% lethal dose (LD_{50}) of certain poison for a kind of animals is known. Now five such animals are injected with this dose. Denoting the number of deaths with X , calculate the probabilities of the events $X = 0, 1, 2, 3, 4$ and 5, respectively.

Solution $\pi = 0.50, n = 5$. With the formula of (2.14),

$$P(0) = \binom{5}{0} (0.5)^5 (0.5)^0 = 0.03125,$$

$$P(1) = \binom{5}{1} (0.5)^4 (0.5)^1 = 0.15625,$$

$$P(2) = \binom{5}{2} (0.5)^3 (0.5)^2 = 0.31250,$$

$$P(3) = \binom{5}{3} (0.5)^2 (0.5)^3 = 0.31250,$$

$$P(4) = \binom{5}{4} (0.5)^1 (0.5)^4 = 0.15625,$$

$$P(5) = \binom{5}{5} (0.5)^0 (0.5)^5 = 0.03125.$$

One can see that although the 50% lethal dose is used, it is still possible to have none of the 5 animals dying after injection and it is also possible to have all the 5 animals dying; the values of probability functions are symmetric about the most possible cases $X = 2$ and $X = 3$ due to that $\pi = 0.50$.

2.3.2 Plot of probability function

Four plots for binomial distributions are shown in Fig. 2.3. One can see that the values of the variable X on the horizontal axis can only be integers $0, 1, 2, \dots$, and the bars are taller around center and shorter on two sides. If the sample size n is not very large, the plot shows positive skew with a long tail along the positive direction of the axis when the parameter $\pi < 0.5$; negative skew with a long tail along the negative direction of the axis when the parameter $\pi > 0.5$; and symmetric when the parameter $\pi = 0.5$ (See Figs. 2.3(a)–(c)). If the sample size n is very large such that both $n\pi$ and $n(1 - \pi)$ are big enough, the plot shows symmetric though the parameter $\pi \neq 0.5$ (see Fig. 2.3(d)).

2.3.3 Population mean and population variance

Let us return to the example of taking balls from a pocket. Each time of taking a ball from the pocket is regarded as one trial; the outcome is denoted by a random variable Y with two possible values, $Y = 1$ for black ball and $Y = 0$ for white ball; the probability of the event $Y = 1$ is π and that of the event $Y = 0$ is $1 - \pi$. Due to (2.5) and (2.6), the population mean and

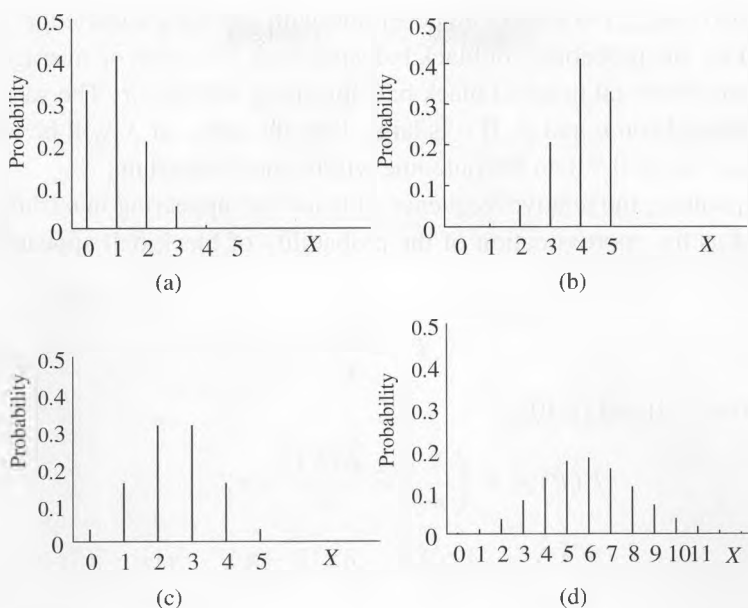


Fig. 2.3 The bar plots for some random variables with binomial distributions (a) $\pi < 0.5$; (b) $\pi > 0.5$; (c) $\pi = 0.5$; (d) $\pi \neq 0.5$, n big enough.

population variance of Y are

$$E(Y) = (1 - \pi) \cdot 0 + \pi \cdot 1 = \pi, \quad (2.16)$$

$$\text{Var}(Y) = (1 - \pi)(0 - \pi)^2 + \pi(1 - \pi)^2 = \pi(1 - \pi). \quad (2.17)$$

Now independently repeat the same trial for n times, of which the outcomes are denoted with Y_1, Y_2, \dots, Y_n respectively and the total number of times the black ball appear is denoted by X ,

$$X = Y_1 + Y_2 + \dots + Y_n.$$

Due to (2.11) and (2.12),

$$\begin{aligned} E(X) &= E(X_1 + X_2 + \dots + X_n) \\ &= E(X_1) + E(X_2) + \dots + E(X_n) = n\pi, \end{aligned} \quad (2.18)$$

$$\begin{aligned} \text{Var}(X) &= \text{Var}(X_1 + X_2 + \dots + X_n) \\ &= \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n) = n\pi(1 - \pi). \end{aligned} \quad (2.19)$$

(2.18) and (2.19) are indeed consistent with our daily knowledge. If for each time, the probability of black ball appearing is π , then on average, out of n times the total times of black ball appearing will be $n\pi$. The variation of X depends on n and π . If n is large, then the range of X will be wider; if π is closer to 0.5 then the outcome will be more uncertain.

In practice, the relative frequency of black ball appearing in n trials (P) is used as the approximation of the probability of black ball appearing in any one trial (π).

$$P = \frac{X}{n}. \quad (2.20)$$

From (2.9) and (2.10),

$$E(P) = E\left(\frac{X}{n}\right) = \frac{E(X)}{n} = \pi, \quad (2.21)$$

$$\text{Var}(P) = \text{Var}\left(\frac{X}{n}\right) = \frac{\text{Var}(X)}{n^2} = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}. \quad (2.22)$$

The above results are not limited to the problem of taking ball from a pocket. In general, for any binomial random variable $X \sim B(\pi, n)$, if the population mean and population variance of X are denoted by μ_x and σ_x^2 , and those of the frequency $P = X/n$ are denoted by μ_p and σ_p^2 , then we have

$$\begin{aligned} \mu_x &= n\pi, \quad \mu_p = \pi, \\ \sigma_x^2 &= n\pi(1-\pi), \quad \sigma_p^2 = \frac{\pi(1-\pi)}{n}, \\ \sigma_x &= \sqrt{n\pi(1-\pi)}, \quad \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}. \end{aligned} \quad (2.23)$$

2.3.4 Discussion on cure rate

Assume the patients meeting certain criteria can be regarded as the individuals of a homogeneous population, the outcomes of a treatment (cure or not) are independent of each other, and the probability of being cured is $\pi = 0.40$ for each. Now randomly select two groups of patients, 30 patients for each, to accept the treatment; as a result, seven patients in group 1 and 14 in group 2 are cured. Why is it that the two groups of patients

with the "same" condition and same treatment gain so different cure rates? Why is it that both of their cure rates 7/30 and 14/30 are not equal to the probability $\pi (= 0.40)$?

Under the above assumptions, to treat 30 patients in group 1 and another 30 in group 2 can be regarded as two independent samples of the binomial population, that is, the number of cured patients in each group is a random variable $X \sim B(0.40, 30)$.

In the first sample, $X = 7$, the cure rate is 7/30. In fact, the probability of such a situation is

$$P(X = 7) = \binom{30}{7} (0.40)^7 (0.60)^{23} = \frac{30!}{7!23!} (0.40)^7 (0.60)^{23} = 0.02634.$$

In the second sample, $X = 14$, the cure rate is 14/30. The probability of such a situation is

$$\begin{aligned} P(X = 14) &= \binom{30}{14} (0.40)^{14} (0.60)^{16} \\ &= \frac{30!}{14!16!} (0.40)^{14} (0.60)^{16} = 0.11013. \end{aligned}$$

When $X = 12$, the cure rate will be $12/30 = 0.40$. The probability of such a situation is

$$\begin{aligned} P(X = 12) &= \binom{30}{12} (0.40)^{12} (0.60)^{18} \\ &= \frac{30!}{12!18!} (0.40)^{12} (0.60)^{18} = 0.14738. \end{aligned}$$

One can see that the situations of $X = 7, 14, 12$ are all possible. $X = 12$ is more likely than others, but $P(X = 12)$ is not equal to 1 so that the observed value in sample is not always equal to the population's cure rate. In fact, the cure rate in sample is a random variable; 7/30 in group 1 and 14/30 in group 2 are the observed values.

The above discussion reminds us that the cure rates in different samples may be quite different. In practice, when two different medicines are taken by two groups of similar patients respectively, even though the two cure rates are quite different, one should not immediately conclude the difference sample between the two population cure rates. Whenever we see a difference between two samples cure rates, it is important to distinguish two possible

conditions: whether it is due to the difference between two samples while the population cure rates are the same; or it is initially due to the difference between two population cure rates.

2.4 Poisson Distribution

2.4.1 Probability function

Poisson distribution was developed by a French mathematician S D Poisson (1781–1840), which can be regarded as the limit of Binomial distribution $B(\pi, n)$ for small π and large n . It is often used to describe the distribution of the number of rare “articles” in a plane or space.

Now take the pulse count of radioactive isotopes as an example. Denote the average number of pulses recorded during a specified period with λ , and divide the period into n sub-intervals such that the average number of pulses during each sub-interval will be λ/n . Assume

- (1) n is big enough such that there is either one pulse or no pulse in each sub-interval and the chance for two or more pulses in any sub-interval is ignorable (large n and $0-1$);
- (2) The probability of the event that the pulse appears in each sub-interval is λ/n (repeated trials and rare event);
- (3) The events of pulse appearing or not in different sub-intervals are statistically independent of one another (independency).

Then the total number X of pulses appearing in the n sub-intervals follows a binomial distribution $B(\lambda/n, n)$ with a probability function

$$P_n(x) = \binom{n}{x} \left[1 - \frac{\lambda}{n} \right]^{n-x} \left[\frac{\lambda}{n} \right]^x$$

which depends on n . It can be proved in mathematics, when $n \rightarrow \infty$, the limit of $P_n(x)$ will be

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad (2.24)$$

where e is a constant approximately equal to 2.718....

Again take the red blood cell count on glass slide as another example. Denote the average number of red blood cells recorded within a specified

glass slide by λ , and divide the slide into n grids such that the average number of red blood cells within each grid will be λ/n . Under three similar assumptions of “large n and 0-1”, “repeat and rare event” and “independency” as mentioned above, one can also obtain that when $n \rightarrow \infty$, the total number X of red blood cells observed on the slide follows a distribution with a probability function given in (2.24).

In general, if the probability function of a random variable X is (2.24), then we say that this variable follows a Poisson distribution with parameter λ , denoted by $X \sim \Pi(\lambda)$.

For many rare diseases (non-infectivity, non-permanent immunity, non-heredity), the number of patients in the population also approximately follows the above three assumptions. Assume the number of individuals n is large enough and any individual is either being attacked or not; the probability of incidence is π , small enough; non-infectivity, non-heredity and independency. Then the number of patients X approximately follows a Poisson distribution $\Pi(\lambda)$, where $\lambda = n\pi$.

Notice that among the above three assumptions “independency” and “repeat” are quite easy to be missed, and in fact, without these two the Poisson distribution will not hold.

For example, for an infectious rare disease, any individual may either be attacked or not. However, people around the patients may have more chance to be attacked than those not around the patients, then the incidence is not independent of one another and the “trials” are not exactly repeated. Although the conditions of “large n ” and “rare event” are met, the number of patients does not follow a Poisson distribution at all.

Again when the bacteria are clustered, if divide the space into many small cubes, then there must tend to have bacteria in the cubes around the one with bacteria. Therefore, it is not independent or repeated and hence the total number of bacteria does not follow a Poisson distribution either.

2.4.2 Plot of probability function

When the parameter π is small and n is large, the probability function of a binomial distribution $B(\pi, n)$ will approximately equal to that of a Poisson distribution $\Pi(\lambda)$, where $\lambda = n\pi$. One can have some feeling about this statement through the following example.

Table 2.3 The probabilities of $X = 0, 1, 2$ chromosome anomalies out of 100 newborns.

X	$P(X)$	
	$B(0.01, 100)$	$\Pi(1)$
0	$\binom{100}{0}(0.99)^{100}(0.01)^0 = 0.3660$	$e^{-1}(1)^0/0! = 0.3679$
1	$\binom{100}{1}(0.99)^{100-1}(0.01)^1 = 0.3697$	$e^{-1}(1)^1/1! = 0.3679$
2	$\binom{100}{2}(0.99)^{100-2}(0.01)^2 = 0.1849$	$e^{-1}(1)^2/2! = 0.1839$

Example 2.4 Assume that the probability of chromosome anomalies in any newborn is 1%. Calculate the probabilities of the events that there are $X = 0, 1$ and 2 newborns with chromosome anomalies out of 100 through two approaches: binomial distribution and Poisson distribution.

Solution See Table 2.3 for the results. On the basis of the relationship between Poisson distribution and binomial distribution, one can imagine that the plot of probability function for Poisson distribution also stands on the integers $0, 1, 2, \dots$ etc. of the horizontal axis, taller around center and shorter on two sides; when $\lambda \leq 5$, it shows positive skew, similar to Fig. 2.1(a); when $\lambda > 5$, it is approximately symmetric, similar to Fig. 2.1(d). In any case, it is impossible to have a plot with negative skew. Why?

2.4.3 Population mean and population variance

We have already known that the population mean and population variance of a binomial distribution are

$$\mu_x = n\pi \quad \text{and} \quad \sigma_x^2 = n\pi(1 - \pi).$$

One can imagine, when $n\pi = \lambda$, $n \rightarrow \infty$ and $\pi \rightarrow 0$, the population mean and population variance of the Poisson distribution $\Pi(\lambda)$ will be

$$\mu_x = \lambda \quad \text{and} \quad \sigma_x^2 = \lambda,$$

where the population mean and population variance are both equal to λ . This is a property of the Poisson distribution specifically, with which people often

identify a Poisson distribution by observing whether the sample mean and sample variance are approximately equal.

2.4.4 Additive property

Another special point of the Poisson distribution is the additive property. If random variables $X_1 \sim \Pi(\lambda_1)$ and $X_2 \sim \Pi(\lambda_2)$ are independent of each other, then

$$X_1 + X_2 \sim \Pi(\lambda_1 + \lambda_2).$$

For instance, if the total pulse count of radioactive isotope recorded in 10 minutes follows a Poisson distribution $\Pi(\lambda)$ and two independently repeated records are made, denoted by X_1 and X_2 , then $X_1 + X_2 \sim \Pi(2\lambda)$.

Note that, if $X \sim \Pi(\lambda)$, then $2X$ does not follow $\Pi(2\lambda)$ and $X/2$ does not follow $\Pi(\lambda/2)$ either.

For instance, by doubling the pulse count of radioactive isotope within 10 minutes, the result $2X$ does not equal to the record made in 20 minutes $X_1 + X_2$; similarly, half of the pulse count of radioactive isotope within 10 minutes does not equal to the record made in 5 minutes either.

2.5 Normal Distribution

2.5.1 Probability density function

In practice, the shape of frequency density histograms of many continuous random variables looks taller around center, shorter on two sides and is symmetric. A family of probability density functions is used

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (2.25)$$

to describe such kind of random variables and say that they follow normal distributions or Gauss distributions, where $\exp(\cdot)$ refers to $e^{(\cdot)}$. It was traditionally regarded that only such kind of distributions is related to normal situations. In fact, it is not true because many probability density functions in real life may not be symmetric and even not taller around center and shorter on two sides. Therefore, now we can only regard "Normal" as a name of a family of distributions, which does not mean a normal situation.

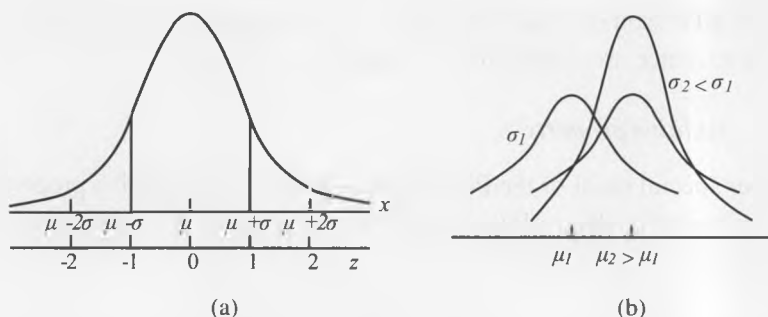


Fig. 2.4 Plots of probability density functions for normal distributions: (a) general situation; (b) changes with μ and σ^2 .

There are two parameters for a normal distribution, μ and σ^2 . μ is the population mean; σ^2 is the population variance (always greater than 0). Any normal distribution is determined by these two parameters so that it can be briefly denoted by $N(\mu, \sigma^2)$. When $\mu = 0, \sigma^2 = 1$, the probability density function and distribution function become

$$\begin{aligned}\varphi(z) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \\ \Phi(z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp\left(-\frac{x^2}{2}\right) dx.\end{aligned}\quad (2.26)$$

Such a normal distribution is called standard normal distribution, denoted by $N(0, 1)$.

From the plots of normal probability density functions (Fig. 2.4), one can learn the following properties intuitively:

- (1) Symmetric about $X = \mu$;
- (2) Peak at $X = \mu$;
- (3) Two inflection points at $X = \mu \pm \sigma$;
- (4) Area under the curve equals 1;
- (5) If σ^2 is fixed, the location of the curve changes with μ so that μ is called location parameter;
- (6) If μ is fixed, the curve is fatter and shorter for bigger σ^2 , thinner and taller for smaller σ^2 so that σ^2 is called shape parameter.

2.5.2 Area under the normal probability density curve

To any normal variable $X \sim N(\mu, \sigma^2)$, after a transformation

$$Z = \frac{X - \mu}{\sigma} \quad (2.27)$$

one will have $Z \sim N(0, 1)$. (2.27) is called standardization transformation, and Z is called standardized normal deviate or Z -value, which in fact is a measure for the deviate from the mean μ with the standard deviation σ as a unit.

A numerical table of the distribution function $\Phi(Z)$ for standard normal distribution is usually attached in most textbooks of statistics. Due to symmetry, given the value of $\Phi(Z)$, the value of $\Phi(-Z)$ can be calculated by

$$\Phi(-Z) = 1 - \Phi(Z)$$

so that this kind of table is often made for $Z \geq 0$ or $Z \leq 0$ only.

With such a table, one can easily obtain the area under the standard normal density curve within any specified interval, that is, the probability of the event that the standard normal variable falls in the interval. For instance, using Table 1 in Appendix II, the areas under the standard normal density curve within intervals $(-1, 1)$, $(-2, 2)$, $(-3, 3)$ are

$$\begin{aligned} \Phi(1) - \Phi(-1) &= \Phi(1) - [1 - \Phi(1)] = 2\Phi(1) - 1 \\ &= 2(0.8413) - 1 = 0.6826, \end{aligned}$$

$$\Phi(2) - \Phi(-2) = 2\Phi(2) - 1 = 2(0.997725) - 1 = 0.99545,$$

$$\Phi(3) - \Phi(-3) = 2\Phi(3) - 1 = 2(0.99865) - 1 = 0.9973.$$

The area under a general normal density curve within any specified interval can be calculated after standardization. For example, the area under the density curve of normal distribution $N(\mu, \sigma^2)$ within the intervals of $(\mu - 1.96\sigma, \mu + 1.96\sigma)$ and $(\mu - 2.58\sigma, \mu + 2.58\sigma)$ are

$$\begin{aligned} &\Phi\left(\frac{(\mu + 1.96\sigma) - \mu}{\sigma}\right) - \Phi\left(\frac{(\mu - 1.96\sigma) - \mu}{\sigma}\right) \\ &= \Phi(1.96) - \Phi(-1.96) = 2\Phi(1.96) - 1 = 2(0.975) - 1 = 0.950, \end{aligned}$$

$$\Phi\left(\frac{(\mu + 2.58\sigma) - \mu}{\sigma}\right) - \Phi\left(\frac{(\mu - 2.58\sigma) - \mu}{\sigma}\right) \\ = \Phi(2.58) - \Phi(-2.58) = 2\Phi(2.58) - 1 = 2(0.995) - 1 = 0.990.$$

The above results show that although any normal variable may take values anywhere in $(-\infty, +\infty)$, the chance that its value falls in $(\mu - 1.96\sigma, \mu + 1.96\sigma)$ is always 95% and the chance that its value falls in $(\mu - 2.58\sigma, \mu + 2.58\sigma)$ is always 99%.

In the applications of normal distribution, the value of a standard normal variable is called two-sided critical value Z_α if the total area under the standard normal density curve within the two tails $|Z| > Z_\alpha$ is equal to α , that is

$$P(|Z| > Z_\alpha) = \alpha.$$

Based on Table 1 in Appendix II, we can have several important critical values in Table 2.4, which will be used everywhere in routine statistical works.

2.5.3 Determination of reference range

In medical field, towards a useful index (a variable in statistics), researchers frequently try to measure a large group of “normal” people to determine the reference range or “normal range” of such an index. If someone’s value is outside this range, then he or she becomes suspect and intensive attention needs to be paid. The group of “normal” people should be a large random sample of the population such that when the frequency density histogram looks like a normal distribution the above knowledge could be used to estimate the 95% or 99% range of the population by cutting two small tails with areas 5% or 1%. However, μ and σ are always unknown so

Table 2.4 Several important critical values for standard normal distribution.

Two-sided Z_α	Area of one tail	Area of two tails
1.645	0.05	0.10
1.960	0.025	0.05
2.576	0.005	0.01

that when the sample size is large enough, they can be replaced by sample mean \bar{x} and sample standard deviation s and then the 95% or 99% range of the index are approximately estimated with $(\bar{x} - 1.96s, \bar{x} + 1.96s)$ and $(\bar{x} - 2.58s, \bar{x} + 2.58s)$.

How large should the sample size be? Although there does not exist any widely recognized criterion yet, by experience, n should be greater than 100 and if \bar{x} and s do not change too much with the increase of sample size, then the sample size might be regarded as appropriate.

One must note that, the 95% reference range just tells that the measures of 95% of "normal" people are within this range; it does not claim that anyone is normal if the measure is in this range; and it does not claim either that anyone is abnormal if the measure is not in this range. Therefore, the reference range could never be a criterion for diagnosis.

The percentile is used to determine the reference range for the index with non-normal distributions or unknown distributions. The 95% inference range of the two-sided index is estimated as the interval of $(P_{2.5}, P_{97.5})$; one-sided 95% inference range is $(-\infty, P_{95})$ and $(P_5, +\infty)$. Since the method of percentile does not adequately use the sample information, it is more reliable to estimate reference range by the method of normal distribution than by percentile if the index follows a normal distribution (or follows a normal distribution after certain transformation of variables).

2.5.4 Normal approximation of binomial distribution and Poisson distribution

2.5.4.1 Correction for continuity

Both binomial distribution and Poisson distribution are distributions for discrete random variables, which can only take values from integers 0, 1, 2, ... In order to borrow distributions of continuous random variables for probability calculation, first of all, the probability function should become "continuous": the "bar" in the plot of probability function being reformed as a "rectangle", that is, for any $X = k$, the bar there is replaced by a rectangle which stands up on the interval $(k - 0.5, k + 0.5)$ with width 1 and the same height as that of the bar (Fig. 2.5(a,b)); obviously, the area of the rectangle and the height of the bar are the same in value, and both are

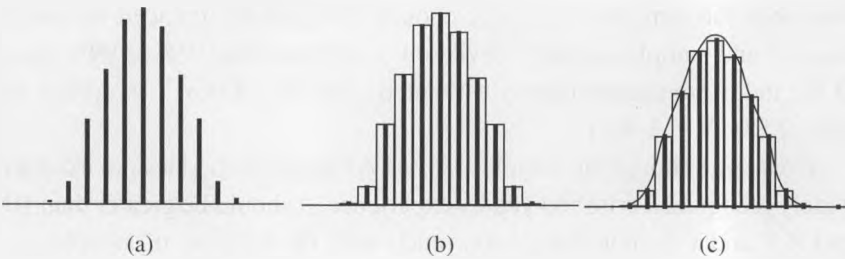


Fig. 2.5 Demonstration for continuity correction and normal approximation of binomial distribution. (a) Bar plot; (b) histogram after continuity correction; (c) normal approximation.

Table 2.5 The continuity correction and normal approximation for probabilities of binomial distribution.

Probability of binomial distribution	Interval for the rectangle standing up	Interval for the area under the approximate normal density curve	Approximate formula for probability: The area under the approximate normal density curve within the corresponding interval
$P(X = k)$	$(k - 0.5, k + 0.5)$	$(k - 0.5, k + 0.5)$	$\Phi\left(\frac{k + 0.5 - n\pi}{\sqrt{n\pi(1 - \pi)}}\right) - \Phi\left(\frac{k - 0.5 - n\pi}{\sqrt{n\pi(1 - \pi)}}\right)$
$P(X \leq k)$	$(0, k + 0.5)$	$(-\infty, k + 0.5)$	$\Phi\left(\frac{k + 0.5 - n\pi}{\sqrt{n\pi(1 - \pi)}}\right)$
$P(X \geq k)$	$(k - 0.5, n)$	$(k - 0.5, +\infty)$	$1 - \Phi\left(\frac{k - 0.5 - n\pi}{\sqrt{n\pi(1 - \pi)}}\right)$
$P(k_1 \leq X \leq k_2)$	$(k_1 - 0.5, k_2 + 0.5)$	$(k_1 - 0.5, k_2 + 0.5)$	$\Phi\left(\frac{k_2 + 0.5 - n\pi}{\sqrt{n\pi(1 - \pi)}}\right) - \Phi\left(\frac{k_1 - 0.5 - n\pi}{\sqrt{n\pi(1 - \pi)}}\right)$

equal to the probability corresponding to $X = k$. This is just the process of correction for continuity. After that, $P(X = k)$, $P(X \leq k)$, $P(X \geq k)$ and $P(k_1 \leq X \leq k_2)$ etc., can be calculated through the areas of the rectangles. See column 2 of Table 2.5.

2.5.4.2 Normal approximation

It can be proved in theory, when n is large such that $n\pi$ and $n(1 - \pi)$ are big enough, the profile of the plot for the probability function of binomial distribution $X \sim B(\pi, n)$ approximates the probability density curve of the normal distribution

$$X \sim N(n\pi, n\pi(1 - \pi))$$

and

$$P = \frac{X}{n} \sim N\left(\pi, \frac{\pi(1 - \pi)}{n}\right).$$

Thus, when n is large such that $n\pi$ and $n(1 - \pi)$ are big enough, after continuity correction the total area of the rectangles in an interval can be replaced by the area under the normal density curve within certain interval. See columns 3 and 4 of Table 2.5.

Similarly, since the Poisson distribution is close to a binomial distribution when n is large and π is small, it can also be approximated by a normal distribution as long as λ is big enough. If $X \sim \Pi(\lambda)$ and λ are big enough, then

$$X \sim N(\lambda, \lambda).$$

Thus, the first three columns of Table 2.5 still hold and the formulas in column 4 need to be changed accordingly, that is, the $n\pi$ in the numerators and $n\pi(1 - \pi)$ in the denominators are all replaced by λ .

Example 2.5 Suppose the probability of detecting a disease by a screening test is 0.005 and randomly selected 10,000 individuals in the population have accepted such a test. What is the probability of the event that 55 individuals with such a disease are detected completely.

Solution Assume that the individuals in this population can be regarded as homogeneous, hence the total detected number can be regarded as a random variable following a binomial distribution. Therefore,

$$P(X \geq 55) = 1 - \sum_{k=0}^{54} \binom{10000}{k} (0.995)^{10000-k} (0.005)^k = 0.2572.$$

Or by Poisson distribution with $\lambda = 10000 \times 0.005 = 50$,

$$P(X \geq 55) = 1 - \sum_{k=0}^{54} \frac{e^{-50} 50^k}{k!} = 0.2577.$$

With the above two approaches one will encounter complicated computations. Now turn to normal approximation,

$$n\pi = 50, \quad n\pi(1 - \pi) = 50 \times 0.995 = 49.75.$$

Using the third formula for binomial distribution in column 4 of Table 2.5,

$$\begin{aligned} P(X \geq 55) &= 1 - \Phi\left(\frac{55 - 0.5 - 50}{\sqrt{49.75}}\right) \\ &= 1 - \Phi(0.638) = 1 - 0.7383 = 0.2616. \end{aligned}$$

Using the formula for Poisson distribution,

$$\begin{aligned} P(X \geq 55) &= 1 - \Phi\left(\frac{55 - 0.5 - 50}{\sqrt{49.75}}\right) \\ &= 1 - \Phi(0.638) = 1 - 0.7383 = 0.2616. \end{aligned}$$

Both are close to 0.2572 (which is obtained by direct calculation) with relative errors less than 2%.

2.5.5 P-P plot and Q-Q plot

Normal distribution is very important to statistic analysis, and it is usually the basis of choosing the appropriate statistical method. As we mentioned in Chap. 1, histogram and stem-and-leaf plot are often used to understand the distribution of the data, but they cannot figure out how closely between the distribution of the data and a normal distribution. However, the P-P plot and Q-Q plot can visually assess whether the data follow a normal distribution.

The P-P plot (Probability-Probability plot) makes a plot with the cumulative distribution of the data versus an assumed cumulative distribution function to assess how closely the assumed distribution fits the data. To perform a test for normality, the cumulated frequencies of the dataset are plotted on the x -axis, and the expected corresponding cumulated frequencies of the standard normal distribution are plotted on the y -axis. If the plot falls on

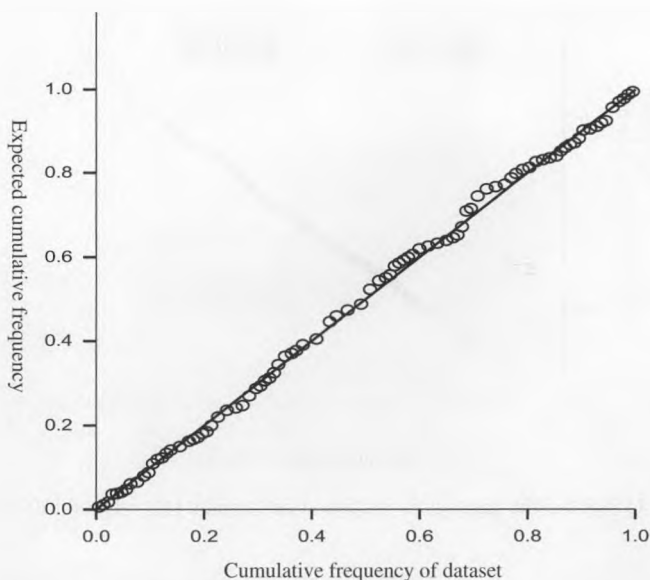


Fig. 2.6 P-P plot for the data set of red blood cell counts ($10^{12}/L$) of 120 normal male adults.

a straight line from (0, 0) to (1, 1), then the data follow a normal distribution. Any deviation indicates a difference between the two distributions. See Fig. 2.6.

The Q-Q plot (Quantile-Quantile plot) is similar with the P-P plot, the only difference is that the Q-Q plot plots the quantile values of the distribution of data against those of the expected distribution. To perform a test for normality, the quantile values of the dataset are plotted on the x -axis, and the corresponding quantile values of the standard normal distribution are plotted on the y -axis. Similar to the P-P plot, if the Q-Q plot falls on a straight line from (0, 0) to (1, 1), then the data follow a normal distribution. See Fig. 2.7.

2.6 Computerized Experiments

Experiment 2.1 Generating random numbers: Generation of random numbers is the basis of random sampling, computer simulation and randomized trial. In general, there are three ways to get random numbers: table

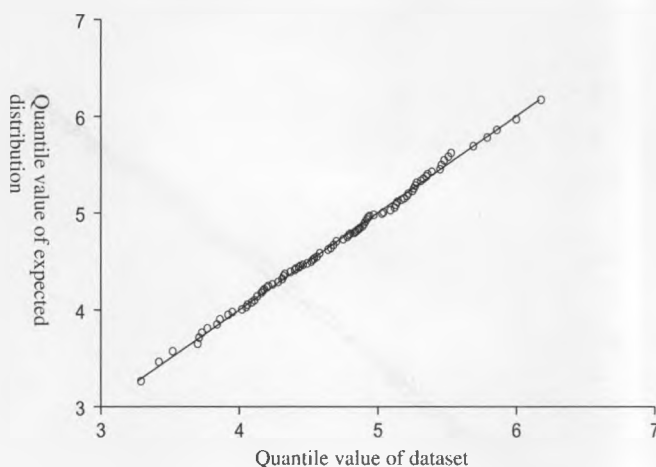


Fig. 2.7 Q-Q plot for the data set of red blood cell counts ($10^{12}/L$) of 120 normal male adults.

for random numbers, certain functional key of calculator and certain function of computer software. Since true random number is inherently not predictable, we usually use computational algorithms to generate relatively random numbers, which are completely determined by an initial value, known as a seed. These relatively random numbers are often called pseudo random numbers. If two sequences of pseudo random numbers are generated by the same seed, the two sequences are the same as well. The content for generating pseudo random numbers (we will omit the word “pseudo” later for convenience) and related testing in SAS are introduced in Program 2.1.

Lines 01–05 generate 100 random numbers from a uniform distribution, where UNIFORM(0) is a random function of the uniform distribution and X takes value between 0 and 1. Lines 06–10 make a plot of the random numbers versus the order of sampling and a frequency histogram to show the independency (or not) and the distribution of the outcomes intuitively.

Discuss this experiment: (1) What should the mean and variance be? (2) How to find the independency among the generated numbers?

Experiment 2.2 Taking balls from a pocket: Assume there are two kinds of balls with the same shape and weight but different colors in a pocket, of which 20% are black balls and 80% are white balls. Each time one ball is taken out of the pocket, record the color and return the ball to the

Program 2.1 Generating random numbers and related testing.

Line	Program	Line	Program
01	DATA RAN;	07	PROC GPLOT;
02	DO I=1 TO 100;	08	PLOT X*I;
03	X=UNIFORM(0);	09	PROC GCHART;
04	OUTPUT;	10	VBAR X/MIDPOINTS=0.05
05	END;		TO 0.95
06	GOPTIONS DEVICE=VGA;	11	RUN;

Program 2.2 Experiment of taking balls from a pocket.

Line	Program	Line	Program
01	DATA BIN;	09	END;
02	DO J=1 TO 10;	10	PROC FREQ;
03	X=0;	11	TABLE X;
04	DO I=1 TO 5;	12	GOPTION DEVICE=VGA
05	Z=UNIFORM(0);	13	PROC CHART;
06	IF Z<=0.2 THEN X=X+1;	14	VBAR X/MIDPOINT=0 TO 5 BY 1;
07	END;	15	RUN;
08	OUTPUT;		

pocket; after stirring, repeat and count the total number of times black ball appears.

Lines 03–07, one trial with $n = 5$ times of taking balls out of the pocket results in a value of X , the total number of times black ball appears out of 5; Z in line 05 is a uniformly distributed random variable. Lines 02–09, 10 trials result in 10 observations of X . Lines 10–15, frequency table and plot.

Discuss this experiment: (1) Pool the results of the whole class to observe the frequency distribution of X and compare with the probability function of the corresponding binomial distribution. (2) Increase n (say $n = 10, 20, 30$) to observe the distribution of X again. (Note: Replace “5” in line 04 with 10, 20, 30 respectively).

Experiment 2.3 P-P plot and Q-Q plot: This is a preparation for plotting P-P plot and Q-Q plot based on Example 1.4. Lines 01–04 read the raw data

Program 2.3 P-P plot and Q-Q plot.

Line	Program	Line	Program
01	DATA RBC;	05	PROC UNIVARIATE;
02	INPUT RBC@@;	06	PROBPLOT RBC/ SQUARE;
03	5.12 5.13 4.58 4.31 4.09	07	RUN;
	4.41 4.33 4.58 4.24 5.45	08	PROC UNIVARIATE;
	...	09	QQPLOT RBC/SQUARE;
04	;	10	RUN;

into SAS. Lines 05–07 plot the P-P plot with the square figure (the default figure is rectangle). Lines 08–10 plot the Q-Q plot with the square figure.

2.7 Practice and Experiments

1. A case-control study on the relationship between esophageal cancer and pickles was performed on all of the patients with esophageal cancer diagnosed in a hospital; and the control group was formed with patients with other acute diseases during the same period in the same hospital, who were matched with the cases according to gender, age and occupation. The individuals in the two groups were interviewed on their intake history of pickles with a standard procedure. The results are summarized in Table 2.6.

Estimate the conditional probabilities

$$P(\text{Frequently taken} \mid \text{Esophageal cancer})$$

and

$$P(\text{Frequently taken} \mid \text{Other acute diseases}).$$

Table 2.6 The data on the relationship between esophageal cancer and pickles.

Pickles	Esophageal cancer	Other acute diseases
Frequently taken	537	554
Not frequently taken	639	922
Total	1176	1476

Can we estimate the conditional probabilities

$$P(\text{Esophageal cancer} | \text{Frequently taken})$$

and

$$P(\text{Other acute diseases} | \text{Frequently taken})?$$

How to evaluate the impact of pickles taken on the incidence of esophageal cancer?

2. According to the statistics, 30% of the patients with acute abdominal pain are suffering from acute appendicitis; 70% of the patients with acute appendicitis have their temperature higher than 37.5°C , while only 40% of the patients with non-acute appendicitis have their temperature higher than 37.5°C . If both “acute abdominal pain” and “temperature higher than 37.5°C ” are taken as the evidence for diagnosis of acute appendicitis, calculate the conditional probability $P(\text{Acute appendicitis} | \text{Acute abdominal pain and temperature higher than } 37.5^{\circ}\text{C})$.

3. Assume the diastolic pressure of healthy high school students follows a normal distribution with mean 9.3 kPa and variance 1.3 kPa. What is the percentage of the students whose diastolic levels are in between 8 kPa and 10.6 kPa, higher than 12.7 kPa and lower than 6.7 kPa respectively?

4. It was required that the missing rate of vaccination among children of certain age should not be higher than 5%. To monitor, 20 randomly selected children were evaluated within each district. It would be ranked as failure if more than one child were missed and excellent if none was missed.

- (1) If the real missing rate of a district is 1%, what is the probability that the district is ranked as failure?
- (2) If the real missing rate of a district is 10%, what is the probability that the district passes the evaluation by a fluke?
- (3) If the real missing rate of a district is 6%, what is the probability that the district is ranked as excellent by a fluke?

5. A board is formed by a net of grids with equal areas. A mass of small particles with equal sizes are scattered randomly over the board and hence

distributed throughout the grids. The possible number of particles in any individual grid may be 0, 1, 2, . . . although the average is 1.2. Estimate:

- (1) What is the percentage of grids not having any particles?
- (2) What is the percentage of grids having some particles?
- (3) What is the percentage of grids having at least four particles?
- (4) What is the percentage of grids having no more than two particles?
- (5) What is the relationship between this problem and the basic principles of the blood cell counting plate?

6. Simulate the process of taking balls from a pocket by computer. Assume there are many balls with the same shape and weight in a pocket, of which 20% are black and 80% are white.

- (1) Sampling 30 times with replacement, record the value for the "total number of times black ball appears";
- (2) Repeating (1) for 200 times and getting 200 values for the "total number of times black ball appears", make a frequency table and a histogram accordingly, then observe whether it is symmetric or not;
- (3) Calculate the sample means and sample variances for the 200 samples in (2) accordingly;
- (4) Calculate the population mean and population variance based on the theory of binomial distribution;
- (5) Compare the results of (3) and (4) and discuss.

7. Think of the differences and similarities between the P-P plot and Q-Q plot.

(1st edn. Jiqian Fang; 2nd edn. Chun Hao, Jiqian Fang)

Chapter 3

Sampling Error and Confidence Interval

For several samples from the same population, usually the sample means are not equal to the population mean and they are different from one another. The difference between sample mean and the population mean are called sampling error. In this chapter, the variation of sampling error will be discussed, and followed by the concept and calculation of confidence interval for the population mean.

3.1 The Distribution of Sample Mean

Let us observe the variation of sample mean through computerized experiment.

3.1.1 *Distribution of sample mean from a population of normal distribution*

Sampling from a normal distribution Assume that the red blood cell counts of healthy males follow a normal distribution $N(4.6602, 0.5746^2)$. Using the program given in Sec. 3.6 of this chapter, 1000 samples with sample size $n = 5$ can be drawn; the sample means of the first 100 samples are showed in the second and sixth columns of Table 3.1. The frequency table and the corresponding histogram are showed in Table 3.2 and Fig. 3.1(a). One can see the following features of sample mean as a random variable:

- (1) Any of the sample means is not necessary equal to the population mean;
- (2) The differences exist among the sample means;
- (3) The distribution of sample means follows certain rule such that more in center, less in two ends and symmetry around the center.

Table 3.1 The sample means, standard errors and 95% confidence intervals of the 100 independent samples with sample size 5, which was randomly drawn from $N(4.6602, 0.5746^2)$ (unit: $10^{12}/L$).

No. of sample	Mean	Standard error	95% confidence interval	No. of sample	Mean	Standard error	95% confidence interval
1	5.00	0.5688	4.2939, 5.7062	51	4.48	0.4006	3.9827, 4.9773
2	4.72	0.3470	4.2891, 5.1509	52	4.32	0.5487	3.6388, 5.0012
3	4.24	0.5763	3.5246, 4.9554	53	4.88	0.3732	4.4167, 5.3434
4	4.64	0.5949	3.9014, 5.3786	54	4.68	0.3524	4.2425, 5.1175
5	4.60	0.4005	4.1028, 5.0972	55	4.80	0.5866	4.0717, 5.5283
6	4.80	0.8186	3.7837, 5.8163	56	4.52	0.3504	4.0850, 4.9550
7	4.68	0.4502	4.1211, 5.2389	57	4.88	0.6869	4.0272, 5.7328
8	4.32	0.8225	3.2989, 5.3411	58	4.80	0.5232	4.1505, 5.4495
9	4.72	0.5964	3.9796, 5.4604	59	4.80	0.2794	4.4531, 5.1469
10	4.40	0.4496	3.8418, 4.9582	60	4.76	0.5823	4.0371, 5.4830
11	4.60	0.5683	3.8944, 5.3056	61	4.76	0.7083	3.8807, 5.6394
12	4.60	0.3401	4.1778, 5.0222	62	4.12	0.5793	3.4008, 4.8392
13	4.60	0.6648	3.7746, 5.4254	63	4.72	0.4419	4.1714, 5.2686
14	4.76	0.6274	3.9811, 5.5389	64	4.44	0.2818	4.0902, 4.7898
15	4.20	0.6886	3.3451, 5.0549	65	4.92	1.0267	3.6454, 6.1947
16	4.64	0.3091	4.2562, 5.0238	66	4.80	0.7191	3.9073, 5.6927
17	4.96	0.4223	4.4357, 5.4843	67	4.72	0.4361	4.1786, 5.2614
18	4.96	0.4083	4.4532, 5.4669	68	4.84	0.5873	4.1109, 5.5691
19	4.68	0.5875	3.9506, 5.4094	69	4.36	0.4892	3.7527, 4.9673
20	4.84	0.5340	4.1771, 5.5030	70	4.76	0.3353	4.3437, 5.1763
21	4.92	0.2852	4.5659, 5.2741	71	4.40	0.4309	3.8650, 4.9350
22	4.60	0.4517	4.0392, 5.1608	72	4.68	0.6880	3.8259, 5.5341
23	4.44	0.4333	3.9021, 4.9779	73	4.60	0.4301	4.0661, 5.1339
24	4.96	0.3711	4.4993, 5.4207	74	4.48	0.6411	3.6841, 5.2759
25	4.64	0.4742	4.0513, 5.2287	75*	4.16	0.3927	3.6724, 4.6476
26	4.96	0.5349	4.2959, 5.6241	76	4.52	0.5487	3.8388, 5.2012
27	4.48	0.4778	3.8868, 5.0732	77	4.36	0.3930	3.8721, 4.8479
28	4.68	0.3818	4.2061, 5.1539	78*	5.04	0.2052	4.7853, 5.2947
29	4.68	0.6289	3.8992, 5.4608	79	4.56	0.9963	3.3231, 5.7969
30	5.28	0.6467	4.4771, 6.0829	80	4.80	0.6243	4.0249, 5.5751
31	4.84	0.6724	4.0053, 5.6747	81*	4.00	0.2090	3.7405, 4.2595
32	4.52	0.3203	4.1224, 4.9176	82	4.64	0.3414	4.2162, 5.0638
33	4.76	0.5841	4.0348, 5.4852	83	5.04	0.4050	4.5372, 5.5428
34	4.48	0.2084	4.2213, 4.7388	84	4.52	0.5353	3.8555, 5.1845
35	5.04	0.6646	4.2149, 5.8651	85	4.44	0.3276	4.0333, 4.8467
36	4.56	0.3912	4.0743, 5.0457	86	4.60	0.3797	4.1287, 5.0713

(Continued)

Table 3.1 (Continued)

No. of sample	Mean	Standard error	95% confidence interval	No. of sample	Mean	Standard error	95% confidence interval
37	4.68	0.5183	4.0366, 5.3234	87	4.48	0.2801	4.1322, 4.8278
38	4.80	0.7445	3.8758, 5.7242	88	4.64	0.2473	4.3330, 4.9471
39	4.72	0.7260	3.8187, 5.6213	89*	5.32	0.3982	4.8256, 5.8144
40	4.68	0.8567	3.6165, 5.7435	90	4.92	0.3473	4.4888, 5.3512
41	4.56	1.0241	3.2887, 5.8313	91	4.72	0.2941	4.3548, 5.0852
42	4.76	0.6786	3.9175, 5.6025	92	4.44	0.4273	3.9096, 4.9704
43	5.04	0.5176	4.3974, 5.6826	93	4.48	0.3594	4.0338, 4.9262
44	4.52	0.3658	4.0659, 4.9741	94	4.92	0.4456	4.3668, 5.4732
45	4.52	0.5944	3.7821, 5.2580	95	4.64	0.4758	4.0494, 5.2306
46	4.72	0.5024	4.0963, 5.3437	96	4.76	0.8516	3.7027, 5.8173
47	5.12	0.6354	4.3312, 5.9088	97	4.64	0.4560	4.0739, 5.2061
48	4.76	0.5837	4.0354, 5.4846	98	4.36	0.3368	3.9419, 4.7781
49*	4.04	0.3595	3.5937, 4.4863	99	4.56	0.6197	3.7907, 5.3293
50	4.52	0.6094	3.7634, 5.2766	100	4.60	0.4566	4.0331, 5.1669

*In fact the confidence interval of this sample did not cover the population mean.

Table 3.2 The frequency table of the sample means, 1000 independent samples with $n = 5$ were drawn from $N(4.6602, 0.5746^2)$.

Lower limit of sub-interval ($10^{12}/L$)	Frequency	Relative frequency (%)	Cumulated relative frequency (%)
3.60–	1	0.1	0.1
3.80–	5	0.5	0.6
4.00–	32	3.2	3.8
4.20–	117	11.7	15.5
4.40–	229	22.9	38.4
4.60–	304	30.4	68.8
4.80–	218	21.8	90.6
5.00–	76	7.6	98.2
5.20–	15	1.5	99.7
5.40–	3	0.3	100.0

- (4) The range of variation for the sample mean is much narrower than that of the initial variable.

One can easily prove in theory: If the random sample with n individuals (X_1, X_2, \dots, X_n) is drawn from a normal distribution $N(\mu, \sigma^2)$,

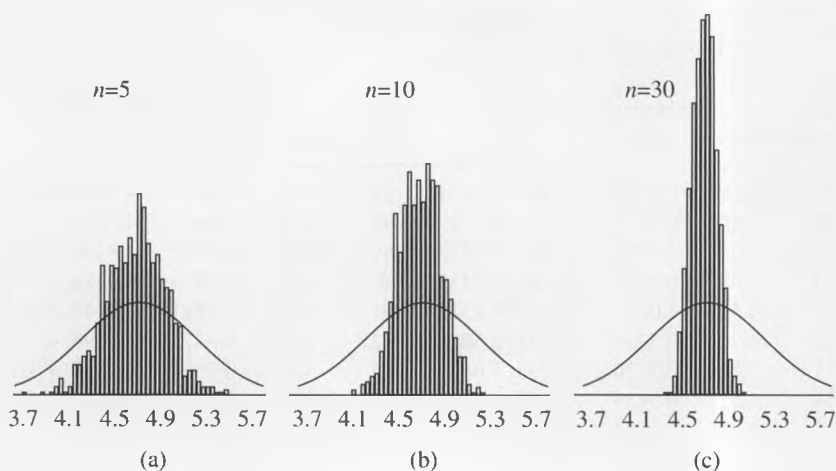


Fig. 3.1 The results of Experiment 3.1. The curve is the density of the initial normal distribution $N(4.6602, 0.5746^2)$; the rest are the histograms of the sample means corresponding to different sample sizes.

then the sample mean, varying from sample to sample, follows a normal distribution

$$\bar{X} \sim N(\mu, \sigma_{\bar{x}}^2). \quad (3.1)$$

The histograms of 1000 independent samples with $n = 5, 10, 30$ respectively are given in Fig. 3.1. It shows:

- (5) The range of variation for the sample means tends to be narrow with the increase of sample sizes.

To distinguish from the standard deviation of the initial variable (σ), in convention, the standard deviation of the sample mean ($\sigma_{\bar{x}}$) is called standard error of sample mean or simply standard error. It is worthwhile to keep in mind that in any case:

$$\text{Standard error of sample mean} = \frac{\text{standard deviation of the population}}{\sqrt{n}},$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}, \quad (3.2)$$

that is, with the cost of n observations, one can only reduce the variation to $1/\sqrt{n}$ times the variation of a single observation.

In practice, the population standard deviation σ is usually unknown, and replaced by sample standard deviation S approximately. Therefore, the estimate of $\sigma_{\bar{x}}$ could be

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}. \quad (3.3)$$

For convenience, $\sigma_{\bar{x}}$ and $S_{\bar{x}}$ can be called theoretical standard error and sample standard error, respectively.

3.1.2 Distribution of sample mean from a population with non-normal distribution

Sampling from a positive skew distribution 1000 independent samples with sample size $n = 5$ were drawn from a positive skew distribution (Fig. 3.2(a)); their sample means were calculated; the corresponding

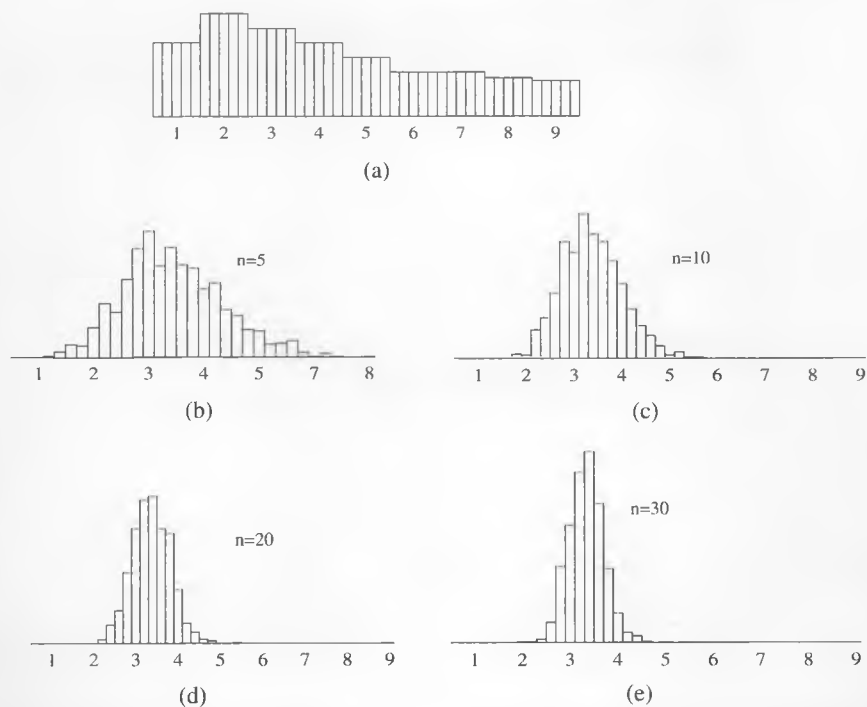


Fig. 3.2 The results of Experiment 3.2. (a) is the initial distribution, positive skew; the rest are the histograms of sample means corresponding to different sample sizes.

histogram is showed in Fig. 3.2(b). Similarly, for sample sizes $n = 10, 20$ and 30 , the corresponding histograms are showed as Fig. 3.2(c), (d) and (e). One can see:

- (1) The distribution of sample means tends to be symmetric with the increase of sample size; when $n = 30$, it looks similar to normal distribution.
- (2) The range of variation for the sample means also tends to be narrow with the increase of sample sizes.

Sampling from an asymmetric hook-like distribution 1000 independent samples with sample size $n = 5$ were drawn from an asymmetric hook-like distribution (Fig. 3.3(a)); their sample means were calculated;

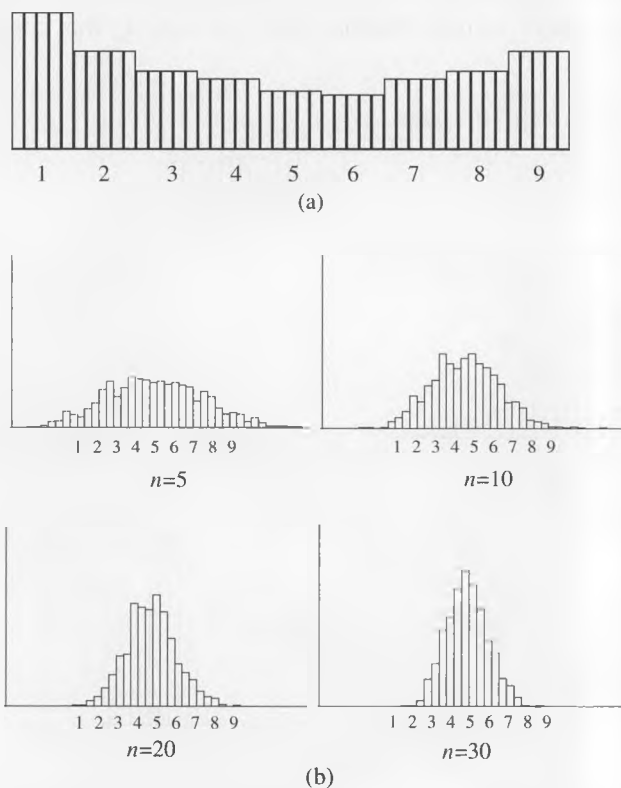


Fig. 3.3 The results of Experiment 3.3. (a) is the initial distribution, hook-like; the rest are the histograms of sample means corresponding to different sample sizes.

the corresponding histogram is showed as Fig. 3.3(b). Similarly, for sample sizes $n = 10, 20, 30$, the corresponding histograms are Fig. 3.3(c), (d) and (e). It is interesting to note that:

- (1) The distribution of sample means is no longer being hook-like; instead, it is quite similar to a normal distribution even when sample size was small.
- (2) The range of variation for the sample means also tends to be narrow with the increase of sample sizes.

The results of Experiments 3.2 and 3.3 reveal a general phenomenon. In fact, it can be proved in theory that: For the population with a non-normal distribution, the distribution of sample means is not a normal distribution though it will be similar to a normal distribution when sample size is big (say, $n \geq 30$); the standard error is still equal to $1/\sqrt{n}$ times the standard deviation for the initial population.

3.2 t Distribution

3.2.1 *Standard normal deviate and standard t deviate*

According to (3.1), the standard normal deviate follows standard normal distribution, denoted as

$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \sim N(0, 1). \quad (3.4)$$

If σ is unknown, S is often used to replace σ and $S_{\bar{X}}$ to replace $\sigma_{\bar{X}}$. Then obviously, $(\bar{X} - \mu)/S_{\bar{X}}$ will no longer follow standard normal distribution. Since $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ is a constant, while $S_{\bar{X}}$ varies with sample, $(\bar{X} - \mu)/S_{\bar{X}}$ must have more variation than $(\bar{X} - \mu)/\sigma_{\bar{X}}$ do. W.S. Gossett (1908) explored the distribution of $(\bar{X} - \mu)/S_{\bar{X}}$, named as t distribution and published under the name of “student”, that is,

$$\frac{\bar{X} - \mu}{S_{\bar{X}}} \sim t \text{ dist.}, \quad \nu = n - 1. \quad (3.5)$$

If $(\bar{X} - \mu)/\sigma_{\bar{X}}$ is called standard normal deviate, then $(\bar{X} - \mu)/S_{\bar{X}}$ could be called standard t deviate. ν in (3.5) is called degrees of freedom of t

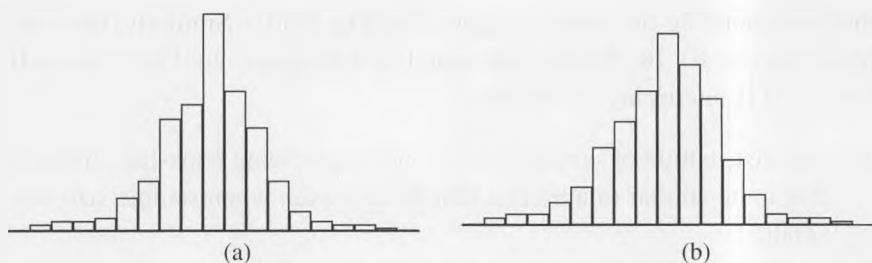


Fig. 3.4 On the basis of 1000 independent samples with $n=5$ drawn from $N(4.6602, 0.5746^2)$ and their sample means, (a) the histogram of the standard normal deviates; (b) the histogram of the standard t deviates.

distribution. Corresponding to different ν , the t distributions differ from each other. In fact, ν is also the degree of freedom of the sample standard deviation S .

3.2.2 The probability density and critical values of t distribution

Standard normal deviate and standard t deviate On the basis of 1000 independent samples with $n = 5$ obtained from Experiment 3.1, the standard normal deviates and standard t deviates were calculated respectively and two histograms can be found in Fig. 3.4.

The two-side probabilities and the corresponding critical values of t distribution are given in Table 5 of Appendix II. For instance, when the degree of freedom is 20, corresponding to two-side probability 0.05, the critical value of t distribution is 2.086, which is greater than 1.96, the two-side critical value of standard normal distribution; corresponding to one-side probability 0.05, the critical value of t distribution is 1.725, which is greater than 1.64, the one-side critical value of standard normal distribution. In general, corresponding to the same probability α , the critical value of t distribution is always greater than that of the standard normal distribution. For instance the, when degree of freedom is 20, corresponding to the critical value of 1.96, the two-side probability is between 0.05 and 0.10, and the one-side probability is between 0.025 and 0.05. In general, corresponding to the same value, the probability α of t distribution is always greater than that of standard normal distribution.

3.3 The Confidence Interval for Population Mean of a Normal Distribution

Assume a population following a normal distribution $N(\mu, \sigma^2)$, where both of μ and σ are unknown. A sample is drawn from this population randomly, of which the mean and standard deviation are denoted as \bar{X} and S respectively.^a The population mean μ is expected to be estimated with an interval.

It has been mentioned that the standard t deviates follows a t distribution. Therefore, 95% of the sample means (but not all) meet the inequality

$$-t_{0.05} \leq \frac{\bar{X} - \mu}{S_{\bar{X}}} \leq t_{0.05}, \quad (3.6)$$

where $t_{0.05}$ is the critical value of t distribution corresponding to the two-side probability 0.05. (3.6) can be rewritten as

$$\bar{X} - t_{0.05}S_{\bar{X}} \leq \mu \leq \bar{X} + t_{0.05}S_{\bar{X}}. \quad (3.7)$$

Assume there is one sample in hand, if it is subject to the above-mentioned “95% of the sample means”, then (3.7) can be used to estimate μ . However, we are not sure whether the sample in hand is subject to the “95% of the sample means”. If we use (3.7) for any sample in hand, and claim μ is located in such an interval, then, in theory, we might be right about 95 times out of 100 times of such way in practice. Therefore, whenever we get the values of $\bar{X} = \bar{x}$, $S = s$ and $S_{\bar{X}} = s_{\bar{x}}$, substitute them into (3.7), get an interval

$$\bar{x} - t_{0.05}s_{\bar{x}} \leq \mu \leq \bar{x} + t_{0.05}s_{\bar{x}} \quad (3.8)$$

we may assume that this is the interval estimate of μ ; however, we have to emphasize that this interval might not necessary cover μ ; the confidence level is just 95%. Hence, (3.8) is called 95% confidence interval of the population mean μ given a random sample of the population.

In general, given a random sample of the population, if the sample size, sample mean and sample standard deviation are denoted as n , \bar{x} and s ; $s_{\bar{x}} = s/\sqrt{n}$; and the t value corresponding to two-side probability α is

^aCapital letters \bar{X} and S are used to indicate sample mean and standard deviation in general, small letter \bar{x} and s are used for the values of a specific sample. Similarly for other occasions, capital letter for variable, small letter for value of the variable.

denoted by t_α , then

$$(\bar{x} - t_\alpha s_{\bar{x}}, \bar{x} + t_\alpha s_{\bar{x}}) \quad (3.9)$$

is called with $(1 - \alpha)$ confidence interval of the population mean μ , $(1 - \alpha)$ is called confidence level. $t_\alpha s_{\bar{x}}$ might be called the precision of the confidence interval, which is the half-length of the interval, indicating the distance between the two ends and the center \bar{x} .

When sample size is big enough, t_α in (3.9) can be replaced by the critical value of standard normal distribution Z_α , that is

$$(\bar{x} - Z_\alpha s_{\bar{x}}, \bar{x} + Z_\alpha s_{\bar{x}}). \quad (3.9a)$$

Confidence interval and confidence level For each sample randomly drawn from the normal distribution before, a confidence interval of μ could be calculated according to (3.9). The 95% confidence intervals of μ for the first 100 samples can be found in the fourth column of Table 3.1. It was easy to find, most (95) intervals had covered the population mean 4.6602, but 5 intervals (Nos. 49, 75, 78, 81 and 89) had not covered it. 95% of the interval estimates were successful and 5% failure. It shows, when we work out an interval estimate based on one set of random sample, the confidence level is about 95%.

Example 3.1 Randomly select 20 cases from the patients with certain kind of disease. The sample mean of blood sedimentation (mm/h) is 9.15, sample standard deviation is 2.13. To estimate the 95% confidence interval and 99% confidence interval of the population mean under the assumption that the blood sedimentation of this kind of disease follows a normal distribution.

Solution

$$\bar{x} = 9.15, \quad s = 2.13, \quad n = 20,$$

$$\bar{x} \pm t_{0.05} s_{\bar{x}} = \bar{x} \pm t_{0.05} \frac{s}{\sqrt{n}} = 9.15 \pm 2.093 \frac{2.13}{\sqrt{20}} = 10.15 \quad \text{and} \quad 8.15,$$

$$\bar{x} \pm t_{0.01} s_{\bar{x}} = \bar{x} \pm t_{0.01} \frac{s}{\sqrt{n}} = 9.15 \pm 2.861 \frac{2.13}{\sqrt{20}} = 10.51 \quad \text{and} \quad 7.78.$$

Hence, of the population mean the 95% confidence interval is (8.15, 10.15), and the 99% confidence interval is (7.78, 10.51).

The above example shows that the confidence interval becomes wider; hence the precision drops when the confidence level is promoted from 95% to 99%. If both of higher confidence level and better precision are expected, then s should be reduced and n should be increased. Usually s is related to the variation among individuals so that it is difficult to be reduced, but increase of sample size is always effective.

3.4 Four Confidence Intervals for Probability and the Difference between Two Probabilities

Confidence intervals for population probability For a random sample of a binomial distribution $B(\pi, n)$, if the number of times a specified event appears is denoted by X , and its frequency is denoted by $P = X/n$, then the population probability π can be estimated through P . When sample size is small, the 95% and 99% confidence interval of π can be obtained from Table 3 of Appendix II; when sample size is big enough, π can be estimated by normal approximation for the distribution of P .

It has been known that, for a n large,

$$P \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right).$$

Since $\pi \approx p$, where p is an observed frequency from a sample in hand. Therefore, approximately

$$P \sim N\left(\pi, \frac{p(1-p)}{n}\right).$$

According to (3.9a), the $(1-\alpha)$ confidence interval of the population probability π can be approximately estimated as

$$\left(p - Z_{\alpha} \sqrt{\frac{p(1-p)}{n}}, p + Z_{\alpha} \sqrt{\frac{p(1-p)}{n}}\right). \quad (3.10)$$

Here π , p and $\sqrt{p(1-p)/n}$ play the role of μ , \bar{x} and s/\sqrt{n} in (3.9a).

Example 3.2 93 patients with similar condition of a hospital were randomly divided into two groups. In the first group, 30 out of 48 cases treated with drug A were cured; in the second group, 20 out of 45 cases treated with

drug B were cured. Calculate the 95% confidence intervals for the recovery probabilities of the two drugs.

Solution

$$\begin{aligned} n_1 &= 48, \quad x_1 = 30; \quad n_2 = 45, \quad x_2 = 20, \\ p_1 &= 30/48 = 0.625; \quad p_2 = 20/45 = 0.444; \\ \sqrt{p_1(1-p_1)/n_1} &= 0.070, \quad \sqrt{p_2(1-p_2)/n_2} = 0.074; \\ p_1 \pm 1.96\sqrt{p_1(1-p_1)/n_1} &= 0.762 \quad \text{and} \quad 0.488 \\ p_2 \pm 1.96\sqrt{p_2(1-p_2)/n_2} &= 0.589 \quad \text{and} \quad 0.299. \end{aligned}$$

Therefore, the 95% confidence interval of π_1 is (0.488, 0.762), and that of π_2 is (0.299, 0.589).

3.5 The Sample Size for Estimation of Confidence Interval

3.5.1 Sample size for confidence interval of the mean of normal population

From (3.7), the width of the confidence interval depends on the confidence level, sample standard deviation and sample size. Inversely, given the confidence level ($1 - \alpha$), the expected width of confidence interval (denoted with δ) and the assumed value of standard deviation (denoted with s), the sample size can be estimated through

$$\delta = t_\alpha \frac{s}{\sqrt{n}}.$$

To solve for n , and replace t_α with Z_α of standard normal distribution approximately, we have

$$n = \left(\frac{Z_\alpha s}{\delta} \right)^2. \quad (3.11)$$

This formula shows that large sample size will be needed if the initial population largely varied (big s), and fine precision (small δ), and high confidence level (small α) are expected.

Example 3.3 It is learnt from a pilot study that the standard deviation of a biochemical index is about ten units. In order to have a 95% confidence interval of the population mean, of which half of the width equals to 2.5 units. What is the sample size needed?

Solution Since $s = 10$, $\delta = 2.5$, $Z_{0.05} \approx 2$, the sample size needed is about

$$n = \left(\frac{Z_{\alpha} s}{\delta} \right)^2 = \left(\frac{2 \times 10}{2.5} \right)^2 = 64.$$

3.5.2 Sample size for confidence interval of the probability of binomial population

From (3.11), the width of the confidence interval depends on the confidence level, sample frequency and sample size. Inversely, given the confidence level ($1 - \alpha$), the width of confidence interval (denoted with δ) and the estimate of frequency (denoted with p), the sample size can be estimated through

$$\delta = Z_{\alpha} \sqrt{\frac{p(1-p)}{n}}.$$

To solve for n ,

$$n = \left(\frac{Z_{\alpha}}{\delta} \right)^2 p(1-p). \quad (3.12)$$

This formula shows that large sample size will be needed if the population probability is close to 0.5 (big variation), fine precision (small δ) and high confidence level (small α) are expected.

Example 3.4 It is learnt from a pilot study that the probability of relapse in one year for a disease is about 10%. Now a survey is planned to further estimate the 95% confidence interval for the probability of relapse in one year, of which the half width is required with 3%. What is the sample size needed?

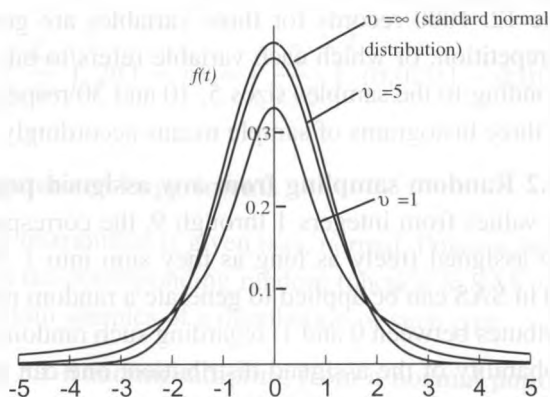


Fig. 3.5 The probability densities for standard normal distribution and t distributions. When $\nu = \infty$, the t distribution tends to standard normal distribution.

Program 3.2 Random sampling from populations in Table 3.3 and distribution of sample means.

Line	Program	Line	Program
01	DATA SAMPLE;	18	END;
02	INPUT K L P1-P8;	19	CARDS;
03	DROP H I J RAN P1-P8 X1-X30 S;	20	1 5 0.15 0.4 0.6 0.75 0.85 0.90 0.95 0.98
04	ARRAY X(30) X1-X30;	21	1 10 0.15 0.4 0.6 0.75 0.85 0.90 0.95 0.98
05	ARRAY P(8) P1-P8;	22	1 20 0.15 0.4 0.6 0.75 0.85 0.90 0.95 0.98
06	DO I=1 TO 1000;	23	1 30 0.15 0.4 0.6 0.75 0.85 0.90 0.95 0.98
07	S=0;	24	2 5 0.25 0.4 0.5 0.58 0.63 0.67 0.75 0.85
08	DO J=1 TO L;	25	2 10 0.25 0.4 0.5 0.58 0.63 0.67 0.75 0.85
09	RAN=UNIFORM(0);	26	2 20 0.25 0.4 0.5 0.58 0.63 0.67 0.75 0.85
10	X(J)=1;	27	2 30 0.25 0.4 0.5 0.58 0.63 0.67 0.75 0.85
11	DO H=1 TO 8;	28	;
12	IF RAN>P(H) THEN X(J)=H+1;	29	GOPTIONS DEVICE=VGA;
13	END;	30	PROC GCHART;
14	S=X(J)+S;	31	VBAR MM/MIDPOINTS=1.2 TO 8.8 BY 0.2;
15	END;	32	BY K L;
16	MM=S/L;	33	RUN;
17	OUTPUT;		

Lines 11–13 are for judgment of the range that the random number falls in and assign the number to X (for instance, 3). Line 16 is for calculation of sample mean. Line 17 is for output. Lines 30–32 are for the plot of histograms of the sample means.

3.7 Practice and Experiments

1. Explain the reason why the following statements hold:

- (1) If out of n independent repeated observations from a binomial distribution the number of times a specified event appears is denoted as $X = x$, and the frequency is denoted as $p = x/n$, then the $(1 - \alpha)$ confidence interval of the population mean of $X (= n\pi)$ can be approximately calculated according to

$$np \pm Z_{\alpha} \sqrt{np(1 - p)}.$$

- (2) If the number of times a specified event appears follows a Poisson distribution, and the observed value $X = x$ is big enough, then the $(1 - \alpha)$ confidence interval of the population mean of X can be approximately calculated according to

$$x \pm Z_{\alpha} \sqrt{x}.$$

2. If ten random samples were prepared from a water source, 1 ml for each, and cultured with plating method under the same condition, the total colony counts was 144, estimate the 95% confidence interval of the colony counts per ml in the water source.
3. There is a uniform die in both the shape and mass, of which 1, 2, 3, 4, 5 and 6 points are painted on the 6 planes respectively. Use such a die or use computer simulation to perform the following experiments: (might collaborate with several students)
 - (1) Independently throw 3000 times, and record the points;
 - (2) For the first 500 in the record of (1), calculate a mean for every 5 values successively such that 100 means are obtained;
 - (3) For the first 1000 in the record of (1), calculate a mean for every 10 values successively such that 100 means are obtained;

- (4) For the first 2000 in the record of (1), calculate a mean for every 20 values successively such that 100 means are obtained;
 - (5) For the first 3000 in the record of (1), calculate a mean for every 30 values successively such that 100 means are obtained;
 - (6) Work out frequency tables and histograms with the data generated in (1), (2), (3), (4), (5) respectively and observe their features.
 - (7) Try to summarize some rule on the basis of the phenomena observed in (6).
4. One can see the formulae of " $\bar{x} \pm s$ ", " $\bar{x} \pm 1.96s$ ", " $\bar{x} \pm 1.96s/\sqrt{n}$ " very often in the literatures. What are they and what are the differences among them?

(1st edn. Jiqian Fang; 2nd edn. Chun Hao, Jiqian Fang)

Chapter 4

Hypothesis Testing for Continuous Variables

The construction of a confidence interval introduced in Chap. 3 is to estimate the range of a population parameters (such as μ and π) with the measures based on sample (such as \bar{X} and p), called interval estimation, which is subject to a kind of statistical inference. This chapter will discuss another kind of statistical inference, called hypothesis testing. Hypothesis testing and interval estimation are not substantially different in principle, but different in the ways of consideration. In practice, hypothesis testing and interval estimation could be used together.

4.1 Specific Logic and Main Steps of Hypothesis Testing

In Example 3.1, 20 cases were randomly selected from the patients with a kind of disease, where the mean and standard deviation of blood sedimentation (mm/h) were 9.15 and 2.13 respectively and we were interested only in the range of population mean. Instead, if we wanted to know whether the population mean was equal to 10.50, then it was one of the typical problems of hypothesis testing.

In fact, the 95% confidence interval (8.15, 10.15) has been given in Example 3.1. It did not cover 10.50, with which one might reasonably exclude the situation that the population mean was equal to 10.50. This showed that we might solve the problem of hypothesis testing with confidence interval.

Let us try another way of thinking: First of all, assume that there is no difference between the population mean and the given constant 10.50; then by analyzing the data, one may judge in what extent the data of current sample support such a hypothesis; finally make a decision, either to accept

or reject the hypothesis of no difference. This is the specific statistical logic of hypothesis testing.

Now let us introduce the main steps of hypothesis testing through the above example.

4.1.1 *Setup the statistical hypotheses*

In Example 3.1, with the current sample mean 9.15, there are two possible situations: one, the population mean has no difference with 10.50, from which the sample mean 9.15 is different from 10.50 due to the sampling error; another, the sample comes from a population whose mean initially differs from 10.50. The two situations are all possible, but only one of the two is correct. Thus, we have to make a choice between two "hypotheses": one is " $H_0 : \mu = 10.50$ ", called null hypothesis; another is " $H_1 : \mu \neq 10.50$ ", called alternative hypothesis.

There is no way to determine which hypothesis is correct and which is not. The feasible way is studying the data to see which hypothesis is more contradicting with the data and reject it. H_0 is relatively simple and clear-cut, under which one can easily find the statistical distribution of the sampling error; while H_1 includes various unknown circumstances, under which one can hardly describe any statistical regulations. Therefore, it is focused on whether the sample information considerably contradicts with H_0 or not. If H_0 does contradict with the sample information, then it is rejected; otherwise, not rejected.

4.1.2 *Select statistics and calculate its current value*

In this example, X follows a normal distribution $N(\mu, \sigma^2)$, but σ^2 is unknown. If $H_0 : \mu = 10.50$ holds, then with the knowledge in Chap. 3, the statistics

$$t = \frac{\bar{X} - 10.50}{S/\sqrt{n}} \sim t \text{ dist.}, \quad \nu = n - 1$$

can be used. The value of t could be large or small; in most occasions it takes value around 0, and sometimes might be far away from 0. This is the regulation followed by the statistics under the hypothesis H_0 (relatively speaking, under the hypothesis H_1 , the regulations followed by the statistics are not so clear and simple).

Putting $\bar{X} = 9.15$, $S = 2.13$ and $n = 20$ into the expression of the statistics t , the current value of it is

$$t = \frac{9.15 - 10.50}{2.13/\sqrt{20}} = -2.8345, \nu = 20 - 1 = 19.$$

4.1.3 Determine the P -value

In order to see whether the current situation ($t = -2.8345$, $\nu = 19$) contradicts with H_0 or not, a so-called P -value corresponding to the current value of t needs to be considered. P -value is defined as a probability of the event that the current situation and even more extreme situation towards H_0 appear in the population.

In our example, the “current situation” corresponds to $\bar{X} = 9.15$ and $t = -2.8345$, and the “even more extreme situation towards H_0 ” corresponds to \bar{X} further away from μ (10.50), hence $t < -2.8345$ and $t > 2.8345$. Therefore,

$$P = P(|t| \geq 2.8345).$$

This is the total area of the two tails within $t < -2.8345$ and $t > 2.8345$ under the probability density curve of the t distribution with degrees of freedom 19. By checking with Table 5 in Appendix II for t distribution, one will find that the P -value is in between 0.01 and 0.02 (Fig. 4.1). Note that the t distribution is symmetric around 0, and the two tails corresponding to $t < -2.8345$ and $t > 2.8345$ are equal.

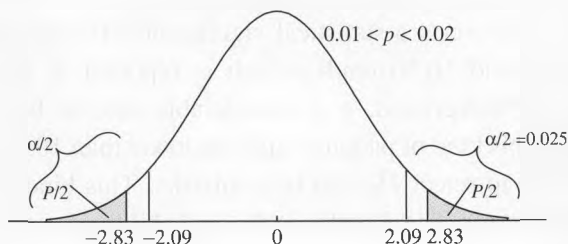


Fig. 4.1 Demonstration for the current value of t and the P -value.

4.1.4 Decision and conclusion

An ignorable small probability α should be defined in advance such as $\alpha = 0.05$ (or 0.01); then the above P -value (in between 0.01 and 0.02) could be regarded as small or almost zero. In other words, under the condition of H_0 , the current situation and even more extreme situation are not quite possible to appear. That is, a small P -value indicates that the information does not support hypothesis H_0 .

However, a “not quite possible situation” actually happens to us in one sample. Then we have two choices: one, still believe that H_0 is true and accept it although we get a “not quite possible situation”; another, believe that a “not quite possible situation” is almost an “impossible situation” and reject it. Since P -value is very small, the second choice is more reasonable.

In general, the decision is: When $P \leq \alpha$, reject H_0 ; otherwise, not reject H_0 .

How to report the result of a hypothesis testing? First of all, give the pre-assigned hypotheses and small probability α as well as the calculated values of statistics and related P -value; then give the conclusion incorporating the background of the problem itself.

For convenience of statement, “reject H_0 ” is often stated as “there is a statistically significant difference” or “the difference is statistically significant”, but it does not mean that the difference is big or obvious; accordingly, “not reject H_0 ” is often stated as “there is no statistically significant difference” or “the difference is not statistically significant”. Note that “not reject H_0 ” means that there is not enough evidence to reject H_0 and it does not straightforwardly mean to “accept H_0 ”. If one likes, it might be understood as to “accept H_0 temporarily” or “the difference is not statistically significant yet”.

The result of the above example might cover: $t = -2.8345$, $P < 0.02$, reject H_0 , that is, there is a statistically significant difference between the population mean and 10.50 mm/h, which is reported in the literatures. Incorporating the background, it is considerable that the blood sedimentation (mm/h) of this kind of patients might be lower than 10.50 on average. It is obvious that, to reject H_0 will be a mistake. This kind of mistake is called with type I error. The P -value is the probability of type I error when H_0 is true.

We have introduced the steps and specific logic of hypothesis testing through an example. More testing procedures will be introduced following such logic in the rest of this chapter and even the whole book.

4.2 The t Test for One Group of Data under Completely Randomized Design

The design is called completely randomized design if the individuals to be observed are completely randomly selected from the population.

Based on the mean and standard deviation of a sample with n individuals randomly selected from a normal distribution $N(\mu, \sigma^2)$, if one wants to judge whether the population mean μ is equal to a given constant μ_0 , the t test for one group of data under completely randomized design can be used. The following are the main steps:

(1) Set up the statistical hypotheses

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0. \quad (4.1)$$

(2) Select statistics and calculate its current value Based on the available knowledge, the statistic t is selected as the test statistic. When H_0 is true,

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t \text{ dist.} \quad (4.2)$$

This kind of statistics is often used for comparison of means, which in fact is to measure the difference between two means with the sample standard error as a unit (called standard t deviate).

(3) Determine the P -value Put the values of \bar{X} , S , n and μ_0 in (4.2) to get the current value of the test statistic t , and check the table of t distribution (degrees of freedom = $n - 1$) in Appendix II to get the two tails accordingly

$$P = P(|t| \geq |\text{current value of statistic } t|). \quad (4.3)$$

(4) Decision and conclusion Comparing the P -value with the pre-assigned small probability α , if $P \leq \alpha$, then reject H_0 ; otherwise, not reject H_0 . Finally, the conclusion can be drawn from the background.

The test used in the example on the blood sedimentation in the last section is exactly the t test for one group of data under completely randomized

design, where the distribution of the variable blood sedimentation (mm/h) is assumed as normality. The test hypotheses are $H_0 : \mu = 10.50$, $H_1 : \mu \neq 10.50$. The alternative hypothesis H_1 includes two sides, $\mu > 10.50$ and $\mu < 10.50$. In general, a test is called a two-side test if its alternative hypothesis includes two sides. When and only when one of the two sides is impossible and hence could be reasonably excluded by the knowledge of subject matter, the alternative hypothesis could be expressed as

$$H_1 : \mu > \mu_0 \text{ or } H_1 : \mu < \mu_0. \quad (4.4)$$

This is called one-side alternative hypothesis. In general, a test is called a one-side test if its alternative hypothesis includes one side only. There is no difference between one-side test and two-side test in terms of their logic and steps, only but the P -value needs to be changed. The P -value for one-side test is the single tail corresponding to the current value of t statistic; when the current value of t statistic is greater than 0,

$$P = P(t \geq \text{current value of statistic } t) \quad (4.5a)$$

when it is less than 0,

$$P = P(t \leq \text{current value of statistic } t). \quad (4.5b)$$

Comparing (4.3) and (4.5a) or (4.5b), one can see, for the same data set, the P -value corresponding to a one-side test will be a half of that corresponding to a two-side test so that H_0 is easier to be rejected. It should be decided at the design stage whether one-side test or two-side test is chosen. Of course, to choose a one-side test, one has to provide adequate reason.

Example 4.1 A large scale survey had reported that the mean of pulses for healthy males is 72 times/min. A physician randomly selected 25 healthy males in a mountainous area and measured their pulses, resulting in a sample mean of 75.2 times/min and a standard deviation of 6.5 times/min. Can one conclude that the mean of pulses for healthy males in the mountainous area is higher than that in the general population?

Solution By experience, the pulses of healthy males follows a normal distribution so that this problem can be analyzed with t test for one group of data under completely randomized design. If we consider that the pulses of healthy males in the mountainous area would never be

lower than that in general area on average, then one-side test is adequate. $H_0 : \mu = 72$, $H_1 : \mu > 72$; $\alpha = 0.05$. $t = 2.69 < 0.005 < P < 0.01$ so that H_0 is rejected and hence we can conclude that the mean of pulses for healthy males in that mountainous area is higher than that in the general population.

4.3 The t Test for Data under Randomized Paired Design

A design is called randomized paired design if similar individuals in terms of several important features are paired and two individuals of any pair are randomly assigned to receive two treatments respectively. For instance, two animals with the same gender and from the same nest could be paired; any specimen could be divided into two parts as a pair; for any individual, before and after treatment could be regarded as a pair; the symmetric parts of any individual's body could be regarded as a pair. The special characteristic of the data under paired design is one-to-one corresponding so that we are concerned with the difference of effects within the pair rather than the effect of each individual.

Comparing the completely randomized design (see Sec. 4.4), the advantage of the paired design is to weaken the interference due to the variation among individuals such that its comparability is better for treatment comparison, especially when the variation among individuals is considerably large.

Example 4.2 The weights (kg) of 12 volunteers were measured before and after a course of treatment with a "new drug" for losing weight. The data is given in Table 4.1. Evaluate the effectiveness of this drug.

Solution Take a glance at the data, they looked like two groups: before treatment and after treatment. As a matter of fact, the effect of the treatment is the difference of weights d , as showed in the last column of Table 4.1. This set of differences can be regarded as a sample of the effect on weight losing. Assume the difference follows a normal distribution, and then a zero mean will indicate that the drug is not effective in weight losing. Thus, the problem turns to a hypothesis testing based on one set of data under completely randomized design on whether the population mean is zero or not.

Table 4.1 The data observed in a study of weight losing.

No.	Weight (kg)		Difference $d = X_1 - X_2$
	Pre-treatment (X_1)	Post-treatment (X_2)	
1	101	100	1
2	131	136	-5
3	131	126	5
4	143	150	-7
5	124	128	-4
6	137	126	11
7	126	116	10
8	95	105	-10
9	90	87	3
10	67	57	10
11	84	74	10
12	101	109	-8
			$\sum d = 16$

(1) Set up the statistical hypotheses Denote the population mean of the variable “difference” d with μ_d

$$H_0 : \mu_d = 0, \quad H_1 : \mu_d \neq 0. \quad (4.6)$$

(2) Select statistics and calculate its current value It is known that when H_0 is true, the statistic

$$t = \frac{\bar{d} - 0}{S_d / \sqrt{n}} \sim t \text{ dist.}, \quad \nu = n - 1, \quad (4.7)$$

where \bar{d} and S_d refer to the mean and standard deviation of the variable “difference”, respectively.

$$\begin{aligned} \bar{d} &= 16/12 = 1.33, \\ S_d^2 &= \left[\sum d_i^2 - \frac{1}{12} \left(\sum d_i \right)^2 \right] / (12 - 1) = 62.6061, \\ S_d &= 7.91. \end{aligned}$$

Substituting into (4.7), get the current value of the statistic

$$t = \frac{1.33}{7.91 / \sqrt{12}} = 0.58, \quad \nu = 12 - 1 = 11. \quad (4.8)$$

(3) Determine the P -value Check up the table for t distribution with degrees of freedom $\nu = 11$, get the area of the two tails beyond 0.58 and -0.58 greater than 0.60, that is, $P > 0.60$.

(4) Decision and conclusion Since the P value is substantially large, H_0 cannot be rejected, and the drug cannot be thought as effectiveness.

Obviously, the decision of “not to reject H_0 ” will be a mistake. The mistake is called type II error if the null hypothesis H_0 is not rejected when H_0 is not really true. Conventionally, the probability of making a type II error is denoted by β , which is not very easy to get accurately. Further discussion on this topic will be found in Chap. 5.

The above-introduced method is applicable to the data analysis for randomized paired design in general. One point needs to note is that we assume the variable “difference” d follows a normal distribution instead of the “weight” of pre-treatment X_1 or post-treatment X_2 respectively.

4.4 The Tests for Comparing Two Means Based on Two Groups of Data under Completely Randomized Design

There are two different situations could be understood as “two groups of data under completely randomized design”: one, the individuals are randomly divided into two groups which correspond to two treatments respectively; another, the two groups are randomly selected from two populations respectively. Comparing to paired design, this design is simple and easy to perform. Especially, this design is often applied when the variation within groups is small.

Example 4.3 For the red blood cell counts of males and females, the sample mean, sample standard deviation and sample size are $\bar{X}_1 = 4.66$, $S_1 = 0.47$, $n_1 = 20$ and $\bar{X}_2 = 4.18$, $S_2 = 0.45$, $n_2 = 15$ respectively. Judge whether the population means of males and females are equal or not.

Solution Assume the red blood cell counts of males and females follow normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively, then the task of data analysis is to test whether the two population means are equal, that is, to test

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2. \quad (4.9)$$

There are two different procedures depending on whether the two variances are equal.

4.4.1 Equal variances

Let us start from the situation that the variances of two populations are equal, $\sigma_1^2 = \sigma_2^2$.

If the two sets of sample mean, standard deviation and sample size are denoted with \bar{X}_1 , S_1 , n_1 and \bar{X}_2 , S_2 , n_2 , then the weighted average of S_1^2 and S_2^2 can be applied to estimate σ^2 ,

$$S_c^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \quad (4.10)$$

or as mentioned in (3.10),

$$S_c^2 = \frac{\sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2}{n_1 + n_2 - 2}. \quad (4.11)$$

When $H_0 : \mu_1 = \mu_2$ holds, it can be proved,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_c^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t \text{ dist.}, \quad v = n_1 + n_2 - 2. \quad (4.12)$$

Substitute the values of \bar{X}_1 , \bar{X}_2 , S_c^2 , n_1 and n_2 into the left-hand side to get the current value of the statistic t ; check up the table for t distribution to get the corresponding P -value; comparing with the pre-specified small probability α , if $P \leq \alpha$, then reject H_0 , otherwise, not reject H_0 .

The 95% confidence interval for the difference of the two population means is (0.16, 0.80), which does not cover the value 0, hence, the difference between two population means might not be zero, that is, $\mu_1 \neq \mu_2$.

In fact, the same question can be solved through a hypothesis test:

(1) Set up the statistical hypotheses

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2.$$

(2) Select statistics and calculate its current value It is reasonable to assume $\sigma_1^2 = \sigma_2^2$ because S_1^2 and S_2^2 are close to each other (see next section for a critical test). The pooled estimate of the population variance

is $S_c^2 = 0.2131$. Substitute the values into the right-hand side of (4.12), and get the current value of the statistic t ,

$$t = \frac{4.66 - 4.18}{\sqrt{0.2131(1/20 + 1/15)}} = 3.04.$$

(3) Determine the P -value Check up the table of t distribution, get the area of two tails, $P < 0.01$.

(4) Decision and conclusion Since $P < 0.01$, reject H_0 . Thus, the difference of sample means of the red blood cell counts between males and females are statistically significant; incorporating the sample means, one may conclude that the population mean of males is higher than that of females.

Comparing the above showed procedure of confidence interval and that of hypothesis testing, one can see that they are not substantially different. The former provides an interval for the difference between population means, without P -value; the latter provides P -value without an interval for the difference. Thus, in practice, most of the statisticians suggest to use both and emphasize that three elements should be mentioned in the report: decision (reject H_0 or not), P -value (confidence level), confidence interval (the estimated range of the parameter). This suggestion is widely applicable to various hypothesis tests that will be introduced in the following text.

4.4.2 Unequal variances

When the variances of two populations are not equal, $\sigma_1^2 \neq \sigma_2^2$, there is no reason to pool the two sample variances as (4.10) and (4.11) will no longer be the approximation of σ_1^2 nor that of σ_2^2 . In such a case, two procedures could be used: non-parametric test based on rank and t' test. The former will be given in Chap. 7. Now let us introduce the t' test.

The statistic used in t' test is

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (4.13)$$

of which the distribution is quite complicated and the P -value is difficult to be determined by checking up a simple table of distribution. Fortunately,

the critical value of t' can be obtained approximately by

$$t'_\alpha \approx \frac{w_1 t_{1\alpha} + w_2 t_{2\alpha}}{w_1 + w_2}, \quad (4.14)$$

where $t_{1\alpha}$ and $t_{2\alpha}$ are the critical values of t distributions with degrees of freedom $n_1 - 1$ and $n_2 - 1$ respectively; and $w_1 = S_1^2/n_1$, $w_2 = S_2^2/n_2$. In other words, the critical value of t' approximately equals to a weighted average of the two critical values of t distributions. Since $\sigma_1^2 \neq \sigma_2^2$, the degrees of freedom cannot be simply pooled and hence the critical value of t' tends to be larger than that of t distribution with degrees of freedom $\nu = n_1 + n_2 - 2$.

When the absolute value of the current value of t' is greater than or equal to the approximate critical value of t' , $P\text{-value} \leq \alpha$, reject H_0 ; otherwise, $P\text{-value} > \alpha$, do not reject H_0 .

Example 4.4 $n_1 = 10$ patients and $n_2 = 20$ healthy people are randomly selected and measured for a biochemical index. The mean and standard deviation of the group of patients are $\bar{X}_1 = 5.05$ and $S_1 = 3.21$, and those of the group of healthy people are $\bar{X}_2 = 2.72$ and $S_2 = 1.52$. Judge whether the two population means are equal or not.

Solution The two population means are denoted by μ_1 and μ_2 .

(1) Set up the statistical hypotheses

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2.$$

(2) Select statistics and calculate its current value First of all, look at the ratio of two variances, $(3.21)^2/(1.52)^2 = 4.46$. By experience, the two variances might be different (see next section for a critical test) so that the statistic t' is selected.

Calculate the current value of the statistic t'

$$t' = \frac{5.05 - 2.72}{\sqrt{(3.21)^2/10 + (1.52)^2/20}} = 2.18.$$

(3) Determine the P -value Corresponding to the degrees of freedom $10 - 1 = 9$ and $20 - 1 = 19$, the two-side critical values of t distribution

are $t_{1\alpha} = 2.26$ and $t_{2\alpha} = 2.09$. Thus the approximate critical value of t' is

$$\begin{aligned} t'_{0.05} &\approx \frac{\frac{(3.21)^2}{10}(2.26) + \frac{(1.52)^2}{20}(2.09)}{\frac{(3.21)^2}{10} + \frac{(1.52)^2}{20}} \\ &= \frac{(1.03041)(2.26) + (0.11552)(2.09)}{1.03041 + 0.11552} = 2.24. \end{aligned}$$

Obviously, $|t'| < t'_{0.05}$, $P > 0.05$.

(4) Decision and conclusion Since $P > 0.05$, do not reject H_0 , and hence there is not enough evidence to conclude for this biochemical index that the average level of patients is significantly different from that of healthy people.

This is a typical example, which deserves to be emphasized. At a first glance on the difference of the two sample means, one might think that the two population means are likely different. However, the two sample standard deviations are quite different from each other that the variation among the patients is much larger than that among normal people (this might be a popular phenomenon). Thus the critical value $t'_{0.05}$ must be larger such that from a large difference between sample means like $5.05 - 2.72 = 2.33$ one can still hardly infer that the two population means are different.

4.5 The F -Test for Equal Variances of Two Groups of Data under Completely Randomized Design

As mentioned in the last section, it is necessary to judge whether the two population variances are equal or not before a test for comparing two means, where we just roughly judged by experience that if the two sample variances were close, then the two population variances were equal, otherwise they were not equal. However, what is the criterion for "close"? A test for equal variances is needed, which also follows the same logic and steps as the test about population means.

For two populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ to infer whether σ_1^2 and σ_2^2 are equal, similar steps should be followed:

(1) Set up the statistical hypotheses

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad H_1 : \sigma_1^2 \neq \sigma_2^2. \quad (4.15)$$

Usually $\alpha = 0.10$ is taken as the small probability.

(2) Select statistics and calculate its current value Denoting the sample variances with S_1^2 and S_2^2 , according to the knowledge from mathematical statistics, we have

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F \text{ dist.}, \quad \nu_1 = n_1 - 1, \quad \nu_2 = n_2 - 1, \quad (4.16)$$

where ν_1 and ν_2 are the two degrees of freedom of the F distribution, and they in fact are the degrees of freedom of S_1^2 and S_2^2 respectively. ν_1 is called numerator degrees of freedom, and ν_2 is called denominator degrees of freedom. When $H_0 : \sigma_1^2 = \sigma_2^2$ holds, (4.16) turns to

$$VR = \frac{S_1^2}{S_2^2} \sim F \text{ dist.}, \quad \nu_1 = n_1 - 1, \quad \nu_2 = n_2 - 1. \quad (4.17)$$

Thus, the variance ratio (VR) can be used as the statistics for the test.

In Appendix II of this book, a table is given for the upper critical value F_α , which is equal to the area of the upper tail of the F distribution. Notice that F distribution is not a symmetric distribution so that the lower critical value F'_α corresponding to the lower tail with area α does not equal to F_α , but

$$F'_{\alpha, \nu_1, \nu_2} = \frac{1}{F_{\alpha, \nu_1, \nu_2}}. \quad (4.18)$$

Due to this relation, the table for the lower critical values is not needed, which can be calculated based on (4.18) easily. See Fig. 4.2.

(3) Determine the P -value The larger variance is always taken as the numerator of the statistic VR for convenience. Thus, to a two-side test,

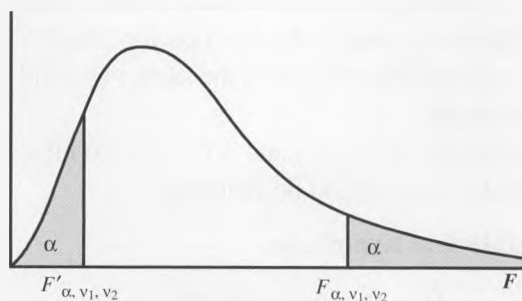


Fig. 4.2 F distribution and its two tails.

given α , one may use $\alpha/2$ to find the upper critical value $F_{\alpha/2}$ of the F distribution, and if the current value of VR is greater than or equal to $F_{\alpha/2}$, then $P \leq \alpha$, otherwise, $P > \alpha$; to a one-side test, given α , one may use it to find the upper critical value of the F distribution F_{α} , and if the current value of VR is greater than or equal to F_{α} , then $P \leq \alpha$, otherwise, $P > \alpha$.

(4) Decision and conclusion If $P \leq \alpha$, then reject H_0 ; otherwise, not reject H_0 . Finally, the conclusion can be drawn from the background.

Returning to Example 4.3, where $VR = S_1^2/S_2^2 = 1.09$, $v_1 = 19$, $v_2 = 14$. Let $\alpha = 0.10$, the two-side critical value of F distribution is $F_{0.10/2} = F_{0.05} = 2.40$. Since $VR < F_{0.05}$, not reject H_0 , hence there is not enough evidence to say that the two population variances are not equal.

Returning to Example 4.4, where $VR = S_1^2/S_2^2 = (3.12)^2/(1.52)^2 = 4.46$, $v_1 = 9$, $v_2 = 19$. Let $\alpha = 0.02$, the two-side critical value of F distribution is $F_{0.02/2} = F_{0.01} = 3.52$. Since $VR \geq F_{0.01}$, reject H_0 , hence one might say that the two population variances are not equal.

In fact, besides serving the test for comparison of two means, this test has its own direct application as follows.

Example 4.5 After stirring, a bottle of test liquid was divided into two groups of 10 species each. The two groups were sent to two laboratories respectively to have their content measured. As a result, the two sample means were equal, but the two sample variances were $S_1^2 = 5$ and $S_2^2 = 3.5$. Judge whether the precisions of measurement in the two laboratories were equal or not.

Solution The problem leads to a test for comparison of two variances, $H_0 : \sigma_1^2 = \sigma_2^2$. $VR = S_1^2/S_2^2 = (5)^2/(3.5)^2 = 2.04$, $v_1 = v_2 = 9$. Let $\alpha = 0.10$, the two-side critical value of F distribution is $F_{0.10/2} = F_{0.05} = 3.18$. Since $VR < 3.18$, $P > 0.10$, not reject H_0 , hence there is not enough evidence to say that the precisions of measurement in the two laboratories were not equal.

Example 4.6 For a measuring procedure used in a multi-center study, the criterion for quality control was defined as that the standard deviation must not be higher than 1.5U. One of the laboratories divided a bottle of test liquid into 10 specimens after stirring, each of which was measured

independently. As a result, the standard deviation was 2.1U. Judge whether the population standard deviation is higher than the criterion for quality control.

Solution This is a special application of the test for equal variance, where the hypotheses are

$$H_0 : \sigma^2 = (1.5)^2, \quad H_1 : \sigma^2 > (1.5)^2.$$

Although here is only one sample, $S_1^2 = (2.1)^2$, $n_1 = 10$, it is wise to regard $(1.5)^2$ as the variance of another sample with sample size ∞ . Thus, $S_2^2 = (1.5)^2$, $n_2 = \infty$. We have $VR = S_1^2/S_2^2 = (2.1)^2/(1.5)^2 = 1.96$, $\nu_1 = 9$, $\nu_2 = \infty$. Let $\alpha = 0.05$, the one-side critical value of F distribution is $F_{0.05} = 1.88$. Since $VR > F_{0.05}$, $P < 0.05$, reject H_0 , hence one might say that the population standard deviation of this laboratory is higher than the criterion for quality control.

4.6 Test for Normality

Normal distribution is one of the most important distributions in statistical analyses, as many statistical methods are based on the assumption that the data follow a normal distribution. There are many methods to test if the data follow normal distribution, such as by graphic interpretation (P-P plot and Q-Q plot) as well as by calculations.

4.6.1 Method of moment

The probability density curve of a normal distribution has certain skewness and kurtosis. Skewness measures whether the density curve is symmetric and a normal density curve has a skewness of zero. Positive skewness indicates that the density curve has a long tail along the right side of the axis (positive skew), while negative skewness denotes the long tail along the left (negative skew). Kurtosis measures the peakedness of the density curve, or the degree of aggregation of the distribution and a normal density curve has a kurtosis of zero. Positive kurtosis indicates that density curve has a high peak and the distribution is centralized, while negative kurtosis indicates that density curve has a flat peak and the distribution is not centralized.

Method of moment is a method eliciting skewness g_1 , kurtosis g_2 , and their standard error σ_{g_1} and σ_{g_2} according to the principles of moments. Using Z test to test the skewness and kurtosis of a distribution, if the results are not statistically significant, the distribution can be regarded as a normal distribution.

The steps of hypothesis testing by using moment method are as follows:

H_0 : The population follows a normal distribution (both zero for skewness and kurtosis of the population);

H_1 : The population does not follow a norm distribution (not both zero for skewness and kurtosis of the population).

We have

$$Z_{g_1} = \frac{g_1}{\sigma_{g_1}}, \quad Z_{g_2} = \frac{g_2}{\sigma_{g_2}}. \quad (4.19)$$

The calculation of g_1 , g_2 , σ_{g_1} and σ_{g_2} are complicated, and it can be completed by using the statistical software. According to the calculated Z values, if both P values of skewness and kurtosis are greater than 0.1 (usually take $\alpha = 0.1$), the distribution can be regarded as a normal distribution.

4.6.2 W test and D test

W test is also known as *Shapiro–Wilk* test. It was developed by S. S. Shapiro and M. B. Wilk to test the normality of data with a not-too-large sample size. D test was developed by D. Agostino to test the normality of data with a large sample size. Both methods are specifically for normality test.

When doing W test or D test, data should be sorted: $x_1 \leq x_2 \leq \cdots \leq x_n$.

In W test, the statistic is the order statistic W :

$$W = \frac{\left\{ \sum_{i=1}^{n/2} a_i [X_{n+1-i} - X_i] \right\}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

X_i is the i th number after the data were sorted, and a_i can be obtained from the table of critical values of W test.

In D test, the statistic Y is:

$$Y = \frac{\sqrt{n}(D - 0.28209479)}{0.02998598}.$$

In the formula above,

$$D = \frac{\sum_{i=1}^n \left(i - \frac{n+1}{2}\right) X_i}{(\sqrt{n})^3 \sqrt{\sum_{i=1}^n [X_i - \bar{X}]^2}},$$

In W test and D test, the null hypotheses are both the data follow a normal distribution. When comparing the statistic W and Y with its critical value, if P is bigger than 0.1, the data being tested follow a norm distribution.

4.6.3 K-S test

Kolmogorov–Simirnov (K-S) is also called K-S one-sample test. It can be used to test whether the data are from a population which follows a specific distribution. In K-S test, cumulative frequency of the observed distribution is compared with that of the tested specific distribution. If the difference between the two sets frequencies is small, the data can be regarded as follow the given distribution. In the test of normality, the null hypothesis of K-S test is also that the population data follow a normal distribution. We have

$$D = \max |T_i - A_i|, \quad (4.20)$$

where T_i denotes the theoretical cumulative frequency of each category of the normal distribution; A_i denotes corresponding sample cumulative frequency. If the null hypothesis holds, the D value of each sampling should not be far away from zero. Hence if the D value is far away from zero (or exceed the critical D value), the null hypothesis should be rejected, and the data cannot be regarded as following a normal distribution.

4.7 The Z-Test for the Parameters of Binomial Distribution and Poisson Distribution (Large Sample)

4.7.1 The Z-test for the population probability of binomial distribution (large n)

It has been mentioned before, if $X \sim B(\pi, n)$, when both $n\pi$ and $n(1 - \pi)$ are large enough, we have approximately

$$X \sim N(n\pi, n\pi(1 - \pi)) \quad (4.21)$$

and

$$P = \frac{X}{n} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right), \quad (4.22)$$

where X is equivalent to the sum of n observations of a 0–1 variable, and $P = X/n$ is equivalent to the sample mean. This property can be used in hypothesis test for population probability of binomial distribution when the sample size is large enough.

4.7.1.1 One sample

Example 4.7 150 physicians being randomly selected from the departments of infectious diseases in a city had received a serological test. As a result, 35 out of 150 were positive. It was known that the positive rate in the general population of the city was 17%. Judge whether the positive rate among the physicians working for the departments of infectious diseases was higher than that in the general population.

Solution Assume the number of physicians with positive result follows a binomial distribution $B(\pi, 150)$.

(1) Set up the statistical hypotheses

$$H_0 : \pi = 0.17, \quad H_1 : \pi > 0.17.$$

When H_0 holds, substituting $\pi = 0.17$ into (4.22), approximately we have

$$P \sim N\left(0.17, \frac{0.17(1-0.17)}{150}\right)$$

or

$$Z = \frac{P - 0.17}{\sqrt{(0.17)(1-0.17)/150}} \sim N(0, 1).$$

(2) Select statistics and calculate its current value The Z is used as statistic, of which the current value is

$$Z = \frac{35/150 - 0.17}{\sqrt{(0.17)(1-0.17)/150}} = 2.06.$$

(3) Determine the P -value Check up the table of standard normal distribution, we have the P -value equal to 0.02.

(4) Decision and conclusion Since $P < 0.05$, reject H_0 , hence one might say that the positive rate among the physicians working for the departments of infectious diseases was higher than that in the general population.

The procedure used for the above example could be extended to general situation. Assume there are n independently repeated trials, and the observed frequency of specified event is p . To judge whether the population probability π is equal to a specified constant π_0 , the following tests can be applied:

(1) Set up the statistical hypotheses

$$H_0 : \pi = \pi_0, \quad H_1 : \pi \neq \pi_0. \quad (4.23)$$

(2) Select statistic and calculate its current value The statistic to be used is

$$Z = \frac{p - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}. \quad (4.24)$$

After substituting the observed frequency p into the above expression, one can get the current value of the statistic.

(3) Determine the P -value By checking up the table of standard normal distribution, the area of the two tails corresponding to the current value of Z and its opposite value will be the P -value accordingly.

(4) Decision and conclusion Comparing the P -value with the pre-specified α , one may make a decision, either reject H_0 or not. And finally, conclusion can be drawn from the background.

When the alternative hypothesis $H_1 : \pi \neq \pi_0$ is changed to a one-side hypothesis $H_1 : \pi > \pi_0$ or $H_1 : \pi < \pi_0$, the statistic will be kept the same, but the one-side P -value should be used.

4.7.1.2 Two samples

Example 4.8 To evaluate the effect of the routine therapy incorporating psychological therapy, the patients with the same disease in a hospital were randomly divided into two groups receiving routine therapy and routine plus

psychological therapy respectively. After a period of treatment, evaluating the same criterion, 48 out of 80 patients in the group with routine therapy were effective, while 55 out of 75 in the other group were effective. Judge whether the probability of effective were different in terms of population.

Solution Assume the two samples were drawn from two binomial distributions $B(\pi_1, n_1)$ and $B(\pi_2, n_2)$ respectively, where $n_1 = 80$ and $n_2 = 75$. Then the problem was led to a hypothesis test.

(1) Set up the statistical hypotheses

$$H_0 : \pi_1 = \pi_2, \quad H_1 : \pi_1 \neq \pi_2.$$

(2) Select statistic and calculate its current value For large samples, from (4.22), the statistics approximately follow distributions

$$P_1 \sim N\left(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1}\right), \quad P_2 \sim N\left(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2}\right).$$

Thus,

$$P_1 - P_2 \sim N\left(0, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right). \quad (4.25)$$

When H_0 holds, the pooled sample frequency is

$$P_0 = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2} = \frac{48 + 55}{80 + 75} = \frac{103}{155}.$$

And it can be used for the approximation of π_1 and π_2 , then

$$P_1 - P_2 \sim N\left(0, \frac{103}{155} \left(1 - \frac{103}{155}\right) \left(\frac{1}{80} + \frac{1}{75}\right)\right)$$

or

$$Z = \frac{P_1 - P_2}{\sqrt{\frac{103}{155} \left(1 - \frac{103}{155}\right) \left(\frac{1}{80} + \frac{1}{75}\right)}} \sim N(0, 1).$$

Taking Z as a test statistic, of which the current value is

$$Z = \frac{\frac{48}{80} - \frac{55}{75}}{\sqrt{\frac{103}{155} \left(1 - \frac{103}{155}\right) \left(\frac{1}{80} + \frac{1}{75}\right)}} = \frac{-0.1333}{\sqrt{0.00575914}} = \frac{-0.1333}{0.0759} = -1.76.$$

(3) Determine the P -value Checking up the table of standard normal distribution, one can get the two-side P -value accordingly, $P = 0.08$.

(4) Decision and conclusion Since $P = 0.08$, do not reject H_0 , hence there is no enough evidence to say that the effects of the two groups are not equal.

The above procedure is easy to be extended to general situation: for the hypotheses (4.25), the statistic used is

$$Z = \frac{P_1 - P_2}{\sqrt{P_0(1 - P_0) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}. \quad (4.26)$$

When the alternative hypothesis $H_1 : \pi_1 \neq \pi_2$ is changed to a one-side hypothesis $H_1 : \pi_1 > \pi_2$ or $H_1 : \pi_1 < \pi_2$, the statistic will be kept the same, but the one-side P -value should be used.

4.7.2 The Z-test for the population mean of Poisson distribution (large λ)

It has been mentioned in Chap. 2 that when λ is large enough, the variable X of Poisson distribution $\Pi(\lambda)$ will approximately follow a normal distribution

$$X \sim N(\lambda, \lambda). \quad (4.27)$$

The hypothesis testing for population mean of Poisson distribution is just based on such a result.

4.7.2.1 Single observation

Example 4.9 The quality control criterion of an instrument specifies that the population mean of radioactivity recorded in a fixed period should not be higher than 50. Now a monitoring test results in a record of 58. Judge whether this instrument is qualified in terms of the population mean.

Solution Assume the radioactivity in the fixed period follows a Poisson distribution $\Pi(\lambda)$.

(1) Set up the statistical hypotheses

$$H_0 : \lambda = 50, \quad H_1 : \lambda > 50.$$

(2) Select statistic and calculate its current value From (4.27), when H_0 is true,

$$X \sim N(50, 50)$$

then the statistic could be used is

$$Z = \frac{X - 50}{\sqrt{50}}$$

and the current value is

$$Z = \frac{58 - 50}{\sqrt{50}} = 1.13.$$

(3) Determine the P -value By checking up the table of standard normal distribution, one can get $P = 0.13$.

(4) Decision and conclusion Since $P > 0.05$, do not reject H_0 , hence there is no enough evidence to say that the instrument does not meet the criterion of quality control.

The above method can be extended to the general situation: Assume a single observation (large enough) is available for the variable X , which follows a Poisson distribution $\Pi(\lambda)$. To test the hypotheses

$$H_0 : \lambda = \lambda_0, \quad H_1 : \lambda \neq \lambda_0 \quad (4.28)$$

(or $H_1 : \lambda > \lambda_0$, or $H_1 : \lambda < \lambda_0$) where λ_0 is a positive constant, when H_0 is true, due to (4.27)

$$\begin{aligned} X &\sim N(\lambda_0, \lambda_0), \\ Z &= \frac{X - \lambda_0}{\sqrt{\lambda_0}} \sim N(0, 1). \end{aligned} \quad (4.29)$$

Substituting the single observation of X into the expression of the statistic, one can get the current value of Z ; then find the P -value and conclude.

The interesting thing is that for a Poisson variable, a single observation is enough for us to estimate a confidence interval (see Chap. 3) and work out a hypothesis testing. This is because the specific characteristic of Poisson distribution only depends on one parameter, which is the population mean as well as population variance.

4.7.2.2 Two observations

Example 4.10 The radioactivity of two specimens was measured for 1 minute independently, resulting in $X_1 = 150$ and $X_2 = 120$ respectively. Judge whether the two corresponding population means in 1 minute are equal or not.

Solution Assume the radioactivity in 1 minute of the two populations all followed Poisson distribution and the two observations were sampled from Poisson distributions $\Pi(\lambda_1)$ and $\Pi(\lambda_2)$ respectively. Then the problem leads to a comparison between the two parameters λ_1 and λ_2 .

In general, one has to go through the following steps as usual to solve this kind of problem:

(1) Set up the statistical hypotheses The problem leads to a test of

$$H_0 : \lambda_1 = \lambda_2, \quad H_1 : \lambda_1 \neq \lambda_2. \quad (4.30)$$

(2) Select statistic and calculate its current value In general, the observations can be regarded as two variables denoted with X_1 and X_2 ,

$$X_1 \sim \Pi(\lambda_1), \quad X_2 \sim \Pi(\lambda_2).$$

When H_0 is true and $\lambda_1 = \lambda_2 = \lambda$ is large enough, from (4.27) approximately we have

$$X_1 \sim N(\lambda, \lambda), \quad X_2 \sim N(\lambda, \lambda).$$

And hence

$$X_1 - X_2 \sim N(0, 2\lambda),$$

$$\frac{X_1 - X_2}{\sqrt{2\lambda}} \sim N(0, 1).$$

By replacing the unknown λ with $(X_1 + X_2)/2$, the test statistic could be used is

$$Z = \frac{X_1 - X_2}{\sqrt{X_1 + X_2}} \sim N(0, 1). \quad (4.31)$$

Substituting the observed values for X_1 and X_2 into the right-hand side of (4.31), one can get the current value of the test statistic.

(3) Determine the P -value By checking up the table of standard normal distribution, one can get the P -value as before.

(4) Decision and conclusion The decision and conclusion can be made after a comparison between the P -value and the pre-assigned α .

Returning to Example 4.10, let us pre-assign $\alpha = 0.05$. The current value of the test statistic is

$$Z = \frac{150 - 120}{\sqrt{150 + 120}} = 1.83.$$

The corresponding two-side P -value is 0.067; since $P > \alpha$, do not reject H_0 , hence the evidence is not enough to say that the two corresponding population means are different.

4.7.2.3 Two "groups" of observations

Example 4.11 The radioactivity of two specimens was independently measured for 10 minutes and 15 minutes respectively, resulting in $X_1 = 1500$ and $X_2 = 1800$. Judge whether the two corresponding population means in 1 minute are equal or not.

Solution Assume the radioactivity measured in 1 minute of the first specimen follows a Poisson distribution $\Pi(\lambda_1)$; the first specimen has been measured for 10 minutes resulting in a total of 1500 and 150/min. on average. Assume that in 1 minute of the second specimen follows a Poisson distribution $\Pi(\lambda_2)$; the second specimen has been measured for 15 minutes resulting in a total of 1800 and 120/min. on average. Then the problem turns to a hypothesis testing of $H_0 : \lambda_1 = \lambda_2$, $H_1 : \lambda_1 \neq \lambda_2$.

In general, assume that there are two Poisson distributed variables defined in the same time unit, $X_1 \sim \Pi(\lambda_1)$ and $X_2 \sim \Pi(\lambda_2)$ respectively, where λ_1 and λ_2 are big enough. Now they are observed for n_1 and n_2 time units respectively. The sum of the records are denoted with $\sum_{i=1}^{n_1} X_{1i}$ and $\sum_{i=1}^{n_2} X_{2i}$. What we are interested in is the hypothesis test of

$$H_0 : \lambda_1 = \lambda_2, \quad H_1 : \lambda_1 \neq \lambda_2. \quad (4.32)$$

From (4.27), we have approximately

$$X_1 \sim N(\lambda_1, \lambda_1), \quad X_2 \sim N(\lambda_2, \lambda_2).$$

The two means are denoted by

$$\bar{X}_1 = \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1}, \quad \bar{X}_2 = \frac{\sum_{i=1}^{n_2} X_{2i}}{n_2}.$$

Obviously, due to the theory of normal distribution, we have approximately

$$\bar{X}_1 \sim N\left(\lambda_1, \frac{\lambda_1}{n_1}\right), \quad \bar{X}_2 \sim N\left(\lambda_2, \frac{\lambda_2}{n_2}\right).$$

And hence,

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\lambda_1 - \lambda_2, \frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}\right). \quad (4.33)$$

When H_0 is true, $\lambda_1 - \lambda_2 = 0$, and replacing λ_1 and λ_2 with \bar{X}_1 and \bar{X}_2 respectively, we have approximately,

$$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \frac{\bar{X}_1}{n_1} + \frac{\bar{X}_2}{n_2}\right). \quad (4.34)$$

And hence

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{\bar{X}_1}{n_1}\right) + \left(\frac{\bar{X}_2}{n_2}\right)}} \sim N(0, 1). \quad (4.35)$$

By substituting the sample data into the right-hand side of the equation, the current value of the test statistic Z and the corresponding P value can be obtained; after comparing with the pre-assigned small probability α , decision about whether to reject H_0 or not can be made.

Back to Example 4.11, by taking 1 minute as a unit of observation, the data of this example can be regarded as the measurements of two groups. That is, the measurement of the first specimen can be regarded as the sum of $n_1 = 10$ times of replication, and that of the second specimen can be regarded as the sum of $n_2 = 15$ times of replication. The problem becomes a hypothesis test of

$$H_0 : \lambda_1 = \lambda_2, \quad H_1 : \lambda_1 \neq \lambda_2.$$

We have

$$\bar{X}_1 = \sum_{i=1}^{10} X_{1i} / 10 = 1500 / 10 = 150,$$

$$\bar{X}_2 = \sum_{i=1}^{15} X_{2i} / 15 = 1800 / 15 = 120.$$

And the current value of the statistic

$$Z = \frac{150 - 120}{\sqrt{(150/10) + (120/15)}} = 6.26.$$

According to the table of standard normal distribution, the P value is almost 0 so that the two population means of the two specimens can be thought as unequal.

4.8 Computerized Experiments

4.8.1 Computer implement of popular hypothesis tests

Experiment 4.1 The t test for data of paired design Find the difference within each pair and then test whether these differences come from a population with mean 0. Demonstrate through Example 4.2.

In Program 4.1, line 2 reads the data pair by pair; line 3 works for difference; line 13 calls for the process MEANS ordering the statistics of mean, standard deviation, standard error, number of cases, t and P -value; line 14 specifies D as the analyzed variable.

Program 4.1 The t test for data of paired design.

Line	Program	Line	Program
01	DATA M_T;	08	101 109
02	INPUT X1 X2;		⋮
03	D=X1-X2;	12	⋮
04	CARDS;	13	PROC MEANS MEAN STDERR N T PRT;
05	101 100	14	VAR D;
06	131 136	15	RUN;
07	... (see Table 4.1)		

Experiment 4.2 The t test for data of two completely random designed groups It includes the t test for equal variances and for unequal variances respectively.

Example 4.12 In a district, 11 patients with acute Ke-shan disease and 13 matched healthy people were recruited and their white phosphorus (mmol/L) was measured. Judge whether the average levels of white phosphorus between the patients with Ke-shan disease and healthy people were different on the basis of the following data:

Patient: 0.84, 1.05, 1.20, 1.20, 1.39, 1.53, 1.67, 1.80, 1.87, 2.07, 2.11

Healthy: 0.54, 0.64, 0.64, 0.75, 0.75, 0.81, 1.16, 1.20, 1.34, 1.35, 1.48, 1.56, 1.87

Program 4.2 The t test for data of two completely random designed groups.

Line	Program	Line	Program
01	DATA GT;	06	;
02	INPUT G X @@;	07	PROC TTEST;
03	CARDS;	08	CLASS G;
04	1 0.84 1 1.05 ... 1 2.11	09	VAR X;
05	2 0.54 2 0.64 ... 2 1.87	10	RUN;

TTEST PROCEDURE

Variable X						
G	N	MEAN	Std	Error	Min	Max
1	11	1.5209	0.4218	0.1272	0.84	2.11
2	13	1.0846	0.4221	0.1171	0.54	1.87

Variances	T	DF	Prob > T
Unequal	2.5239	21.4	0.0196
Equal	2.5237	22.0	0.0193

For H_0 : Variances are equal, $F' = 1.00$, $DF = (12, 10)$, $\text{Prob} > F' = 1.0000$.

The last line is for the test of equal variance. When $\text{Prob} > F'$ is larger than α , the results corresponding to "Equal" are adopted; otherwise, those corresponding to "Unequal" are adopted.

4.8.2 Experiment on two types of error

Experiment 4.3 Type I error Two samples are drawn randomly from the same population; the t test for two sample means is performed; if $P \leq \alpha = 0.05$, then reject H_0 and count as type I error.

In Program 4.3, lines 03–06 generate a sample with 10 individuals; line 4 generates a value of the random variable following $N(2, 1)$; lines 02–07 perform replication; lines 08–10 work for t test.

Program 4.3 Experiment on type I error.

Line	Program	Line	Program
01	DATA ER1;	07	END;
02	DO J=1 TO 2;	08	PROC TTEST;
03	DO I=1 TO 10;	09	CLASS J;
04	X=RANNOR(0)+2;	10	VAR X;
05	OUTPUT;	11	RUN;
06	END;		

Discussion Each student performs five times of this experiment and count the times of type I error. After collecting all the counts of the whole class, estimate the probability of type I error and discuss the reason.

The type II error Two samples with sizes $n_1 = n_2 = 10$ are drawn randomly from the populations $N(2, 1)$ and $N(4, 1)$ respectively; the t test for two sample means is performed; if $P > \alpha = 0.05$, then not reject H_0 and count as type II error.

In Program 4.4, lines 02–06 of draw from the first group of sample from $N(2, 1)$, $n_1 = 10$; lines 07–11 draw from the second group of sample from $N(4, 1)$, $n_2 = 10$; lines 12–14 perform a t test for the two groups. Each

Program 4.4 Experiment on type II error.

Line	Program	Line	Program
01	DATA ER2;	09	X=RANNOR(0)+4;
02	DO I=1 TO 10;	10	OUTPUT;
03	J=1;	11	END;
04	X=RANNOR(0)+2;	12	PROC TTEST;
05	OUTPUT;	13	CLASS J;
06	END;	14	VAR X;
07	DO I=1 TO 10;	15	RUN;
08	J=2;		

student repeats five times and counts the number of times that H_0 is not rejected.

Discussion Dividing the sum of the counts of all the students that H_0 is not rejected by the total number of replications, one can get a percentage, which is the estimate of the probability of type II error; increase the sample sizes $n_1 = n_2 = n$ to see the relationship between the probability of type II error and the sample size n .

4.8.3 Test of normality

Experiment 4.4 To test the normality of white phosphorus value in Example 4.12. In Program 4.5, lines 01–05 read data; line 06 calls for the process PROC UNIVARIATE to do statistical description, and calls NORMAL to do normality test.

Program 4.5 Test of normality.

Line	Program	Line	Program
01	DATA NORM;	05	;
02	INPUT X @@;	06	PROC UNIVARIATE NORMAL;
03	CARDS;	07	VAR X;
04	0.54 2 0.64 . . . 2 1.87	08	RUN;

Results:

Tests for Normality				
Test	Statistic		<i>p</i>	Value
Shapiro–Wilk	W	0.926209	Pr < W	0.3038
Kolmogorov–Smirnov	D	0.202948	Pr > D	0.1471
Cramer–von Mises	W-Sq	0.072543	Pr > W-Sq	0.2422
Anderson–Darling	A-Sq	0.429717	Pr > A-Sq	>0.2500

In the results, four methods were adopted to test the normality of the data. In SAS, when the sample size is not more than 2000, we use results of Shapiro–Wilk; when the sample size is over 2000, we use Kolmogorov–Smirnov. Different statistical software may have different requirements of sample size. If *P* value is over 0.1, we accept the hypothesis that data follow a normal distribution.

4.9 Practice and Experiments

1. To study the effect of whole body microwave exposure (2450MHz) to tumor growth in rats with breast cancer planted, the experimental results were given in Table 4.2:

Table 4.2 The volume of tumors observed at different time points after tumor planting with and without whole body microwave exposure ($\bar{X} \pm S \text{ cm}^3$).

		Number of rats	Days after tumor planting			
			18	25	32	39
10	Exposure	8	0.58 ± 0.60	2.78 ± 2.52	9.42 ± 6.26	21.97 ± 12.41
	Pseudo-exposure	8	0.56 ± 0.38	2.27 ± 2.06	9.27 ± 7.51	20.43 ± 14.61
20	Exposure	8	1.32 ± 1.22	6.82 ± 4.80	20.0 ± 15.3	46.50 ± 32.50
	Pseudo-exposure	6	1.61 ± 1.07	7.56 ± 4.80	23.0 ± 11.7	46.60 ± 25.60
40	Exposure	8	0.47 ± 0.37	7.36 ± 5.70	16.86 ± 9.52	37.90 ± 23.00
	Pseudo-exposure	8	0.60 ± 0.49	7.48 ± 5.06	17.1 ± 10.4	32.80 ± 14.50

Assume the volume of tumors under the same condition follows a normal distribution approximately. Analyze this data set with the knowledge learnt in this chapter.

2. To study the protective effect of cobra toxin to oleic-acid-type respiratory distress syndrome, 76 mice were randomly divided into two groups. The control group (40 mice), caudal vein injection with oleic acid (0.07 ml/kg) in 20–30 minutes after intraperitoneal injection of physiological saline (200 μ l/kg); the experimental group (36 mice), caudal vein injection with oleic acid (0.07 ml/kg) in 20–30 minutes after intraperitoneal injection of cobra toxin (200 μ l/kg). Part of the mice were killed at the end of 1 hour after the injection and the rest were killed at the end of 2 hours after the injection. The measured data were recorded in Table 4.3 ($\bar{X} \pm S$):

Table 4.3 The data on the effect of cobra toxin to oleic-acid-type respiratory distress syndrome.

Killing time	Group	Weight (g)	Wet weight of lung (mg)	Dry weight of lung (mg)	Lung coefficient	Wet/Dry	Water content of lung (%)
1 hr.	Control (27)	21.8 \pm 2.4	397 \pm 83	63 \pm 11	18 \pm 4	6.4 \pm 1.1	84 \pm 4
	Experiment (22)	21.9 \pm 2.5	337 \pm 63	54 \pm 10	15 \pm 3	6.3 \pm 0.8	84 \pm 4
2 hr.	Control (13)	22 \pm 3	414 \pm 62	60 \pm 11	19 \pm 3	7.0 \pm 1.3	85.3 \pm 2.8
	Experiment (14)	21 \pm 4	340 \pm 90	57 \pm 11	16 \pm 4	6.0 \pm 1.0	82.7 \pm 2.7

Assume the variable under the same condition follows a normal distribution approximately. Analyze this data set with the knowledge learnt in this chapter.

3. Since the new therapy is adopted, the ratio between the number of patients being cured and the total number of patients receiving this therapy (sample cure rate) has increased this year. The hypothesis test shows $P < 0.05$. Which of the following inferences are correct?

- (1) If the probability of cure in this year is really equal to that in last year, then the probability that the sample cure rate of this year is higher or even much higher than last year is less than 5%.

- (2) The probability that the type I error happens is less than 5%.
- (3) The probability that the null hypothesis " H_0 : The probability of cure in this year is equal to that in last year" holds is less than 5%.
- (4) The probability that the statement "the above null hypothesis is false" does not hold is less than 5%.
- (5) $P < 0.05$ indicates that the improvement of the probability of cure is statistically significant so that the new therapy is worthwhile to be widely used in clinic.
- (6) The statistical test shows that at least the new therapy is not worse than the old therapy.
- (7) A small P value does not necessary mean a great improvement of the probability of cure. In fact, as long as there is some improvement, even though very little, one can always make the P value smaller by increasing the sample size.
- (8) $P = 0.05$ is not small enough, the smaller the P value, the more the clinic significance.

4. 24 volunteers were recruited for a research project on reducing cholesterol. They were completely randomly divided into two groups with 12 individuals each. Group A received a special diet and group B received a

Table 4.4 The cholesterol (mmol/L) records before and after the study.

Group A			Group B		
No.	Pre-study	Post-study	No.	Pre-study	Post-study
1	6.11	6.00	1	6.90	6.93
2	6.81	6.83	2	6.40	6.35
3	6.48	6.49	3	6.48	6.41
4	7.59	7.28	4	7.00	7.10
5	6.42	6.30	5	6.53	6.41
6	6.94	6.64	6	6.70	6.68
7	9.17	8.42	7	9.10	9.05
8	7.33	7.00	8	7.31	6.83
9	6.94	6.58	9	6.96	6.91
10	7.67	7.22	10	6.81	6.73
11	8.15	6.57	11	8.16	7.65
12	6.60	6.17	12	6.98	6.52

medical therapy. The cholesterol (mmol/L) of each individual was measured before and after the study. The data were showed in Table 4.4.

- (1) Judge whether the pre-study cholesterol level for the two treatments are equal on average.
- (2) Judge whether the two treatments are effective on average respectively.
- (3) Judge whether the effects on reducing cholesterol are equal on average.

5. 90 patients with diabetes were recruited as volunteers involved in a research project, of which 50 received routine treatment and 40 received a new medicine. After a therapeutic period, 15 out of the 50 with routine treatment and 18 out of the 40 with new medicine had their quality of life improved respectively. Judge whether the probabilities of improving quality of life were equal for the patients with the two treatments.

6. The specified length of time in Example 4.10 was 1 minute. If it was regarded as $n_1 = n_2 = 60$ seconds and 1 second was taken as the time unit, apply the test statistic for two groups of observations to judge whether the two population means were equal or not. And comparing with the procedure given in the text do you see any difference and connection? Why both equations (4.31) and (4.35) are suitable for Example 4.10, while Eq. (4.31) is not suitable for Example 4.11?

7. The anxiety levels of 45 randomly selected lying-in women were measured. As a result, the variance of the scores among 25 with their education above senior high school was 125, and the variance of the scores among 20 with their education around senior high school or below was 220. Assuming the two groups of lying-in women were balanced on other aspects, can we conclude that the variation of anxiety level among low educated lying-in women is relatively higher?

8. Practice the following computerized experiments:

- (1) Randomly draw 100 groups of sample from a normal distribution $N(0, 1)$ with sample size 10 for each. Given $\alpha = 0.05$, test $H_0 : \mu = 0$, $H_1 : \mu \neq 0$ for the 100 groups respectively and count the number of times that H_0 is rejected.
- (2) Randomly draw 100 groups of sample from a normal distribution $N(1, 1)$ with sample size 10 for each. Given $\alpha = 0.05$, test $H_0 : \mu = 0$,

$H_1 : \mu \neq 0$ for the 100 groups respectively and count the number of the times that H_0 is not rejected.

- (3) Keeping all the same as (1) except changing α to 0.01, how is the result different from that in (1)?
- (4) Keeping all the same as (2) except changing α to 0.01, how is the result different from that in (2)?

(1st edn. Jiqian Fang; 2nd edn. Jing Gu, Jiqian Fang)



Chapter 5

Chi-Square Test for Categorical Variable

The Z test and t test introduced before are used for continuous variable. Chi-square test for categorical variable will be discussed in this chapter.

5.1 Chi-Square Distribution and Pearson's Goodness-of-Fit Test

The basic theory of chi-square test for categorical variable is a distribution for continuous variable — Chi-square distribution and the goodness-of-fit test for categorical variable.

5.1.1 Chi-square distribution

Assume a variable Z distributes as a standard normal distribution with values ranging from $-\infty$ to $+\infty$. The distribution of Z^2 is different, with values ranging from 0 to $+\infty$, high probability for values close to 0 and low probability for values apart from 0. The curve of its probability density looks like the curve with $\nu = 1$ in Fig. 5.1. The distribution of Z^2 is called χ^2 distribution with degrees of freedom 1, denoted by $\chi^2_{(1)}$. χ^2 is called as chi-square.

Assume there are k independent variables, all follow standard normal distribution, denoted as Z_1, Z_2, \dots, Z_k , then the distribution of their squared sum $Z_1^2 + Z_2^2 + \dots + Z_k^2$ is called a chi-square distribution with $\nu = k$ degrees of freedom, denoted as $\chi^2_{(k)}$. Figure 5.1 gives the probability density curves of $\chi^2_{(3)}$ and $\chi^2_{(5)}$. They have a skewed peak and the skew is improved when the degrees of freedom increases. When ν is very big, the χ^2 distribution closes to a normal distribution and the population mean of χ^2 distribution equals its degrees of freedom.

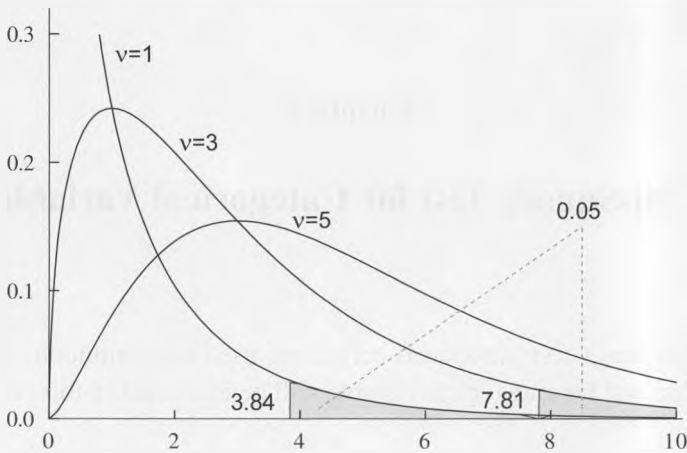


Fig. 5.1 Probability density curves of several χ^2 distributions.

In the Appendix, there is a table for the critical values of chi-square distribution with different degrees of freedom, where the area of upper-tail is α . When degrees of freedom ν is beyond the table, a normal approximated equation may be used to estimate the critical value.

$$Z = \sqrt{2\chi^2} - \sqrt{2\nu - 1} \quad (5.1)$$

or

$$\chi^2 = \frac{1}{2}[Z^2 + 2Z\sqrt{2\nu - 1} + 2\nu - 1]. \quad (5.2)$$

For example, when $\nu = 100$, given $\alpha = 0.05$, the critical value of $\chi_{0.05(100)}^2$ may be estimated by substituting $Z_{0.05} = 1.96$ into Eq. (5.2),

$$\chi_{0.05}^2 = \frac{1}{2}[1.96^2 + 2(1.96)\sqrt{200 - 1} + 200 - 1] = 129.07.$$

This value is very close to the value $\chi_{0.05(100)}^2 = 124.34$ picked from the table in the Appendix.

Suppose $\chi_{(\nu_1)}^2$ and $\chi_{(\nu_2)}^2$ are two independent variables following χ^2 distributions respectively, $\nu_1 > \nu_2$. It can be proved that $\chi_{(\nu_1)}^2 + \chi_{(\nu_2)}^2$ still follows a chi-square distribution with $\nu_1 + \nu_2$ degrees of freedom. Similarly, $\chi_{(\nu_1)}^2 - \chi_{(\nu_2)}^2$ also follows a χ^2 distribution with $\nu_1 - \nu_2$ degrees of freedom.

5.1.2 The χ^2 test for goodness-of-fit (large sample)

One important usage of χ^2 distribution is to test whether a sample comes from a given theoretical distribution or not.

H_0 : The sample comes from a given theoretical distribution.

H_1 : The sample does not come from the theoretical distribution.

Denote the observed frequencies and the expected frequencies (when H_0 is true) with f_i and e_i respectively, $i = 1, 2, \dots, k$. $(f_i - e_i)$ reflects the difference between the actual frequency and theoretical frequency. This difference is a non-continuous random variable, but K. Pearson (1899) proved that when H_0 is true and the sample size is large enough,

$$\chi_P^2 = \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i} \sim \chi^2 \text{ distribution.} \quad (5.3a)$$

It is called Pearson's χ^2 statistic by convention. In Eq. (5.3a), squaring $(f_i - e_i)$ is equal to deal with the positive difference and negative difference; dividing by e_i is to evaluate a relative difference; using e_i as the denominator instead of f_i is to make the statistic more robust.

If the parameters of the theoretical distribution are unknown, in order to get e_i , these parameters should be estimated based on sample data firstly. Therefore, the degrees of freedom of the χ^2 statistic will deducted accordingly, that is

$$\nu = k - 1 \text{ — number of parameters used in estimating } f_i\text{'s.} \quad (5.4)$$

In theory, large sample means an infinite sample size; but in practice, large sample only means a big enough expected frequency. By experience, the expected frequency should not be less than 5.

The above introduced χ^2 test for goodness-of-fit has no restriction on the theoretical distribution thus it has been widely applied. The applications of χ^2 test in different situations will be introduced in the following sections.

Another statistic called the likelihood ratio χ^2 statistic is often used accompanied with the Pearson's χ_P^2 . When H_0 is true, it is known as,

$$\chi_L^2 = 2 \sum_{i=1}^k f_i \ln \left(\frac{f_i}{e_i} \right) \sim \chi^2 \text{ distribution.} \quad (5.3b)$$

Its degree of freedom is the same as that showed in (5.4).

Table 5.1a Data of a binary variable from two independent samples.

	Binary variable		Total
	+	−	
Sample 1	f_{11}	f_{12}	n_{r1} (fixed)
Sample 2	f_{21}	f_{22}	n_{r2} (fixed)
Total	n_{c1}	n_{c2}	n

Table 5.1b Probability expression of the data in Table 5.1a.

	Binary variable		Total
	+	−	
Sample 1	π_1	$1 - \pi_1$	1
Sample 2	π_2	$1 - \pi_2$	1

5.2 The χ^2 Test for Comparison between Two Independent Sample Proportions

The data of comparing the average levels of a binary response variable between two independent samples may be expressed as Table 5.1a. Suppose there are n_{r1} individuals in sample 1 and n_{r2} individuals in sample 2, being observed independently. For instance, n_{r1} patients receive a new medication and n_{r2} patients receive a routine one.

In this kind of data, the variable is binary (+ & −); the comparison is between two independent samples (sample 1 & sample 2); the basic figures $f_{11}, f_{12}, f_{21}, f_{22}$ are listed in a table with four cells. Therefore, it is usually called as data of 2×2 contingency table or fourfold table.

In Chap. 4, we used to apply the method of normal approximation to solve the above problem, that is, to test $H_0: \pi_1 = \pi_2$ by statistic

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p) \left(\frac{1}{n_{r1}} + \frac{1}{n_{r2}} \right)}}, \quad (5.5)$$

where π_1 and π_2 are the population probabilities corresponding to the two samples; and

$$p_1 = \frac{f_{11}}{n_{r1}}, \quad p_2 = \frac{f_{21}}{n_{r2}}, \quad p = \frac{n_{c1}}{n}$$

When H_0 is true and the sample size is big enough, Z will approximately follow the standard normal distribution; hence, Z^2 will approximately follow a χ^2 distribution with 1 degree of freedom. Now let us use the above-mentioned χ^2 test to solve the same problem.

Example 5.1 Before a clinical trial, 215 patients with pulmonary heart disease in a hospital were randomly divided into two groups, of which 164 patients in group 1 took digitalis and 51 patients in group 2 did not take it. Each of them received an ECG examination before the trial starting. The results are listed in Table 5.2. The arrhythmia rate in group 1 is 49.38% and the arrhythmia rate in group 2 is 37.25%. Now the question is whether the two groups of patients can be regarded as "balance in disease condition", or whether the difference of arrhythmia rates in two groups is of statistical significance.

5.2.1 Setting up the testing hypotheses

Same as before, it is to test

$H_0: \pi_1 = \pi_2$ (Two population probabilities are equal)

$H_1: \pi_1 \neq \pi_2$ (Two population probabilities are not equal)

Table 5.2 Data of patients of pulmonary heart disease with arrhythmia.

	ECG		Total	Arrhythmia rate (%)
	Arrhythmia	Normal		
With digitalis	81(76.28)	83(87.72)	164	49.39
Without digitalis	19(23.72)	32(27.28)	51	37.25
Total	100	115	215	46.51

5.2.2 Calculating the current value of statistic

5.2.2.1 The expect frequencies

When H_0 is true, denote $\pi_1 = \pi_2 = \pi$, the combined estimate of the population probability π will be

$$\pi \approx p = \frac{n_{c1}}{n}.$$

Then the expected frequencies will be

$$\begin{aligned} e_{11} &= n_{r1}\pi \approx \frac{n_{r1}n_{c1}}{n}, & e_{12} &= n_{r1}(1 - \pi) \approx \frac{n_{r1}n_{c2}}{n}, \\ e_{21} &= n_{r2}\pi \approx \frac{n_{r2}n_{c1}}{n}, & e_{22} &= n_{r2}(1 - \pi) \approx \frac{n_{r2}n_{c2}}{n}. \end{aligned}$$

That is,

$$e_{ij} \approx \frac{n_{ri}n_{cj}}{n}, \quad i, j = 1, 2, \quad (5.6)$$

where i and j represent the numbers of row and column respectively. For example

$$e_{11} \approx \frac{n_{r1}n_{c1}}{n} = \frac{164 \times 100}{215} = 76.28.$$

The figures listed within parentheses in Table 5.2 are the expected frequencies. It should be noted that the sum of expected frequencies in each row or column must equal to the sum of actual frequencies accordingly.

5.2.2.2 Calculating the value of statistic χ_p^2

(1) Basic formula of Pearson's χ_p^2 and the special case for 2×2 table. For 2×2 table, (5.3a) can be written as

$$\chi_p^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(f_{ij} - e_{ij})^2}{e_{ij}}. \quad (5.7a)$$

For example, for Example 5.1, by (5.7a)

$$\begin{aligned} \chi_p^2 &= \frac{(81 - 76.28)^2}{76.28} + \frac{(83 - 87.72)^2}{87.72} + \frac{(19 - 23.72)^2}{23.72} + \frac{(32 - 27.28)^2}{27.28} \\ &= 2.3028. \end{aligned}$$

(5.7a) can be further simplified as

$$\chi_p^2 = \frac{(f_{11}f_{22} - f_{12}f_{21})^2 n}{n_{r1}n_{r2}n_{c1}n_{c2}} \quad (5.8a)$$

It can be easily proved that formulas (5.5), (5.7a) and (5.8a) are equivalent to each other.

For Example 5.1, by (5.5),

$$Z = \frac{81/164 - 19/51}{\sqrt{(100/215)(115/215)(1/164 + 1/51)}} = 1.5175,$$

$$Z^2 = 2.3028.$$

By (5.8a),

$$\chi_P^2 = \frac{(81 \times 32 - 83 \times 19)^2 \times 215}{164 \times 51 \times 100 \times 115} = 2.3028.$$

(2) Correction for continuity. When $n \geq 40$, if $1 \leq e_{ij} < 5$, the value calculated by formula (5.7a) and (5.8a) will be larger than the value it ought to be, then a correction is needed. It is called the correction for continuity. Corresponding to (5.7a), the correction formula is

$$\chi_P^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(|f_{ij} - e_{ij}| - 0.5)^2}{e_{ij}} \quad (5.7b)$$

Corresponding to (5.8a), it is

$$\chi_P^2 = \frac{(|f_{11}f_{22} - f_{12}f_{21}| - n/2)^2 n}{n_{r1}n_{r2}n_{c1}n_{c2}} \quad (5.8b)$$

After correction, the value of χ_P^2 is smaller than uncorrected one.

In fact, in such a case, certain correction should also be used for the statistic Z : Assume $p_1 > p_2$, substituting

$$p_1 = \frac{f_{11} + 0.5}{n_{r1}}, \quad p_2 = \frac{f_{21} - 0.5}{n_{r2}}$$

into formula (5.5), we may get corrected statistic Z , and the result is equivalent to the correction in (5.7b) and (5.8b).

However, when $n < 40$ or one of e_{ij} happens to be less than 1, the above correction is not enough and then the method of exact probability will be more proper (see Sec. 5.8).

(3) Calculation of likelihood ratio statistic χ_L^2 . For a 2×2 table, when $H_0: \pi_1 = \pi_2$ is true, the formula (5.3b) can be written as

$$\chi_L^2 = 2 \sum_{i=1}^2 \sum_{j=1}^2 f_{ij} \ln \left(\frac{f_{ij}}{e_{ij}} \right). \quad (5.9)$$

For Example 5.1, the χ_L^2 is

$$\begin{aligned} \chi_L^2 = 2 \left[81 \ln \left(\frac{76.28}{81} \right) + 83 \ln \left(\frac{87.72}{83} \right) + 19 \ln \left(\frac{23.72}{19} \right) \right. \\ \left. + 32 \ln \left(\frac{27.28}{32} \right) \right] = 2.3277. \end{aligned}$$

It is different from $\chi_P^2 = 2.3028$. In theory, when the sample size $n \rightarrow \infty$, both χ_P^2 and χ_L^2 tend to the same value. In practice, χ_P^2 is preferred and χ_L^2 as reference.

5.2.3 Calculating degrees of freedom

In order to estimate the expected frequencies e_{ij} , we need to know the values of parameters π_1 and π_2 , but they are unknown. As showed before, we used marginal sums n_{r1} , n_{r2} , n_{c1} and n_{c2} to estimate e_{ij} .

However, when n is fixed, only one of n_{r1} and n_{r2} is independent, and only one of n_{c1} and n_{c2} is independent so that the number of parameters used to estimate e_{ij} is indeed 2. Based on Eq. (5.4), $\nu = 4 - 1 - 2 = 1$.

In general, for a contingency table, we may calculate the degrees of freedom by the following formula

$$\nu = (\text{number of rows} - 1)(\text{number of columns} - 1). \quad (5.10a)$$

For 2×2 table,

$$\nu = (2 - 1)(2 - 1). \quad (5.10b)$$

5.2.4 Determine the P value and conclude

Same as other hypothesis test, given α , one can have the critical value of χ^2 distribution χ_a^2 , if the value of the statistic $\chi^2 > \chi_a^2$, then reject the null hypothesis, $p < \alpha$; or alternatively, having the p value corresponding to the value of χ^2 , if $p < \alpha$, then reject the null hypothesis. Here the statistic χ^2 could either be χ_P^2 or χ_L^2 .

For Example 5.1, $\chi_P^2 = 2.3028$, $\nu = 1$, $\chi_{0.05}^2 = 3.84$, $p > 0.05$ so that H_0 cannot be rejected. That is, we are not able to say that the positive rates of the two groups are statistically different. Based on the statistic $\chi_L^2 = 2.3277$, the conclusion is the same. This makes us more confident to the decision.

Example 5.2 A hypothesis has been proposed that breast cancer in women is caused in part by events that occur between the age at menarche (the age when menstruation begins) and the age at first childbirth. In particular, the hypothesis is that the risk of breast cancer increases as the length of this interval increases. If this theory is correct, then an important risk factor for breast cancer is the age at first birth. This theory would explain in part why breast-cancer incidence seems higher for women in the upper socioeconomic groups, because they tend to have child relatively late.

An international study was set up to test this. Breast-cancer cases were identified among women in selected hospitals in the United States, Yugoslavia, Greece, Brazil, Taiwan, and Japan. Controls were chosen from women of comparable age who were in the hospitals at the same time but who did not have breast cancer. All women were asked about their age at first birth.

The set of women with at least one birth was arbitrarily divided into two categories: (1) women whose age at first birth was ≤ 29 , and (2) women whose age at first birth was ≥ 30 . As Table 5.3 showed the results found among women with at least one birth: 683 out of 3220 (21.2%) women with breast cancer (case women) and 1498 out of 10245 (14.6%) women without breast cancer (control cases) had an age at first birth ≥ 30 . How can we assess whether this difference is significant or simply due to chance? (cited from: *Fundamentals of Biostatistics*)

(1) Setting up the testing hypotheses

$H_0: \pi_1 = \pi_2$ (distributions of age at first birth in two groups are equal)

Table 5.3 Age at first birth for case group and control group.

Group	Age at first birth		Total
	≥ 30	≤ 29	
Case	683	2,537	3,220
Control	1,498	8,747	10,245
Total	2,181	11,284	13,465

$H_1: \pi_1 \neq \pi_2$ (distributions of age at first birth in two groups are not equal).

- (2) **Calculating the current value of statistic** For 2×2 cross tables, we could use the special formula (5.8a).

$$\chi_P^2 = \frac{(683 \times 8747 - 2537 \times 1498)^2 \times 13465}{(683 + 2537)(1498 + 8747)(683 + 1498)(2537 + 8747)} = 78.37.$$

- (3) **Determine the P value and make conclusion** Same as before, the degree of freedom for 2×2 cross tables is 1.

For this example, $\chi^2 = 78.37$, $\chi_{0.05}^2 = 3.84$, $P < 0.05$, so that H_0 can be rejected. That is to say, the distribution of age at first birth in two groups are not equal. According to the sample rates 21.2% and 14.6%, we could judge that age at first birth ≥ 30 was an important risk factor for breast cancer.

Besides, we could use odds ratio (OR) as an index to quantify this kind of risk:

$$OR = \frac{683/2537}{1498/8747} = \frac{683 \times 8747}{2537 \times 1498} = 1.57.$$

That is to say, the risk of occurrence of breast cancer for women whose age at first birth was ≥ 30 was 1.57 times as much as those women whose age at first birth was ≤ 29 .

Example 5.3 The 169 peptic ulcer patients with similar condition were randomly divided into two groups, treated with two drugs losec and

Table 5.4 The treatment effect after four weeks.

Treatment	Effect		Total
	Heal	Not heal	
Losec	64	21	85
Ranitidine	51	33	84
Total	115	54	169

ranitidine respectively. The treatment effect after four weeks is listed in Table 5.4. Now the question is whether the difference of healing rates in the two groups is of statistical significance.

(1) Setting up the testing hypotheses

$H_0: \pi_1 = \pi_2$ (the healing rates of peptic ulcer in two groups are equal)

$H_1: \pi_1 \neq \pi_2$ (the healing rates of peptic ulcer in two groups are not equal).

(2) Calculating the current value of statistic

$$\begin{aligned}\chi_P^2 &= \frac{(64 - 57.84)^2}{57.84} + \frac{(21 - 27.16)^2}{27.16} + \frac{(51 - 57.16)^2}{57.16} \\ &\quad + \frac{(33 - 26.84)^2}{26.84} = 4.13.\end{aligned}$$

(3) Determine the P value and make conclusion For this example, $\chi^2 = 4.13$, $\chi_{0.05}^2 = 3.84$, $P < 0.05$ so that H_0 can be rejected. That is to say, the healing rates of peptic ulcer in two groups are not equal. Because the healing rates were 75.29% and 60.71%, we could judge that losec has a higher healing rate than ranitidine.

(4) Determine the confidence interval for the difference between the two probabilities In order to further describe the difference between the effect of two drugs, we can calculate the confidence interval for the difference between the two probabilities. Let the two sample rates be p_1 and p_2 , if $n_1 p_1$, $n_1(1 - p_1)$ and $n_2 p_2$, $n_2(1 - p_2)$ are all greater than 5, the sampling distribution for the difference between the two sample rates is close to a normal distribution. We can use the normal approximation to calculate the confidence interval for the difference

between the two probabilities:

$$[(p_1 - p_2) - Z_{\alpha/2} S_{p_1-p_2}, (p_1 - p_2) + Z_{\alpha/2} S_{p_1-p_2}], \quad (5.11)$$

where $S_{p_1-p_2}$ is the standard deviation, calculated as follows:

(1) When $\pi_1 \neq \pi_2$, the estimated standard deviation is:

$$S_{p_1-p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

(2) When $\pi_1 = \pi_2$, the estimated standard deviation is:

$$S_{p_1-p_2} = \sqrt{p_c(1-p_c) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

According to (5.11), the 95% confidence interval for the difference between the healing rates of losec and ranitidine is equal to (0.0068, 0.2848).

The difference between the two rates is a percentage. To describe the clinical meaning more clearly, people prefer to calculate the inverse of the difference:

$$(75.29\% - 60.71\%)^{-1} = (14.58\%)^{-1} = 6.86.$$

The meaning is: To have 1 more patient to be cured, about 6.86 patients are needed to shift from the losec group to ranitidine group.

In general, the inverse of the difference between the two rates is called the number of cases need to treat (NNT), which is an indicator comparing the clinical efficacy of two treatments when the outcome is measured by a rate. Obviously, the smaller the NNT, the better the clinical efficacy.

5.3 The χ^2 Tests for Binary Variable under a Paired Design

A single sample with sample size n can be cross classified into four groups according to two binary variables. The data may be listed in the format of Table 5.5a, which is also called a 2×2 contingency table, where f_{ij} is the number of cases belonging to the level i of variable A and level j of variable B . For such kind of design, depending on the study purpose,

Table 5.5a 2×2 cross classified data.

Variable A	Variable B		Total
	1	2	
1	f_{11}	f_{12}	n_{r1}
2	f_{21}	f_{22}	n_{r2}
Total	n_{c1}	n_{c2}	$n(\text{fixed})$

Table 5.5b Probability expression of data in Table 5.5a.

Variable A	Variable B		Total
	1	2	
1	π_{11}	π_{12}	π_{r1}
2	π_{21}	π_{22}	π_{r2}
Total	π_{c1}	π_{c2}	1.0

two kinds of χ^2 tests may be used for testing the independence (between variables A and B) and the difference of two proportions (between variables A and B) respectively.

5.3.1 The χ^2 test for independence between two binary variables

If the probability distribution of variable A is independent to that of variable B , we say that these two variables are independent to each other; otherwise, there is an association between the two variables.

According to the theory of probability, the term "independence between two variables" means that the probability of the joint event equals to the product of two marginal probabilities. Corresponding to Table 5.5a we can have Table 5.5b.

(1) Setting up the testing hypotheses

H_0 : Variable A depends on variable B

H_1 : Variable A is associated with variable B .

When H_0 is true, we will use the same statistics as before:

$$\chi_P^2 = \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i} \sim \chi^2 \text{ distribution} \quad (5.3a)$$

$$\chi_L^2 = 2 \sum_{i=1}^k f_i \ln \left(\frac{f_i}{e_i} \right) \sim \chi^2 \text{ distribution.} \quad (5.3b)$$

(2) **Calculating the current value of statistic** When H_0 is true,

$$\pi_{ij} = \pi_{ri}\pi_{cj}, \quad i, j = 1, 2.$$

By this, we can have the expected frequencies

$$e_{ij} = n\pi_{ij}, \quad i, j = 1, 2.$$

However, we do not know the real values of π_{ij} , which depend on π_{ri} and π_{cj} . Now π_{ri} and π_{cj} can be estimated by the sample data

$$\pi_{r1} \approx \frac{n_{r1}}{n}, \quad \pi_{r2} \approx \frac{n_{r2}}{n}, \quad \pi_{c1} \approx \frac{n_{c1}}{n}, \quad \pi_{c2} \approx \frac{n_{c2}}{n}.$$

And then

$$\begin{aligned} e_{11} = n\pi_{11} &\approx \frac{n_{r1}n_{c1}}{n}, & e_{12} = n\pi_{12} &\approx \frac{n_{r1}n_{c2}}{n}, \\ e_{21} = n\pi_{21} &\approx \frac{n_{r2}n_{c1}}{n}, & e_{22} = n\pi_{22} &\approx \frac{n_{r2}n_{c2}}{n}. \end{aligned}$$

That is,

$$e_{ij} = n\pi_{ij} \approx \frac{n_{ri}n_{cj}}{n}, \quad i, j = 1, 2.$$

This formula is exactly the same as (5.6) before. In fact, the formula for the degrees of freedom is also the same as (5.10b).

It is seen that the χ^2 test for independence between two binary variables based on the data of cross classified 2×2 table has the same formula for calculation as those for the comparison between two independent sample proportions. However, the purpose and design of study as well as the explanation of results are different indeed.

Table 5.6 The results of two immunological tests.

A	B		Total
	+	-	
+	172	8	180
-	12	68	80
Total	184	76	260

If the null hypothesis is rejected, one may further calculate a Pearson's contingency coefficient r_P to describe the strength of the association between the two variables quantitatively. It is defined by

$$r_P = \sqrt{\frac{\chi_P^2}{n + \chi_P^2}} \quad (5.12)$$

r_P takes value between 0 and 1; $r_P = 0$ means totally independence and $r_P = 1$ means a complete association.

Example 5.4 There were 260 serum samples. Each sample was divided into two and tested by two different methods of immunological test of rheumatoid factor respectively. The results are listed in Table 5.6. Now are the results of the two methods independent.

Solution The test hypotheses are

H_0 : A is independent to B, H_1 : A is associated with B

Calculate e_{ij} by (5.6)

$$e_{11} \approx \frac{n_{r1}n_{c1}}{n} = \frac{184 \times 180}{260} = 127.38.$$

Use the same way to get $e_{12} = 52.62$, $e_{21} = 56.62$, $e_{22} = 23.38$.

By formula (5.7a) or (5.8a), $\chi_P^2 = 173.74$.

The degrees of freedom $\nu = 1$, $\chi_{0.05}^2 = 3.84$ so that P is much less than 0.05. H_0 is rejected and the results of the two methods might be associated

to each other. Furthermore,

$$r = \sqrt{\frac{173.74}{260 + 173.74}} = 0.63.$$

The results suggest a positive association between the results of two immunological tests. Since $172 \times 68 - 8 \times 12 > 0$, it means that a positive result of A might likely be related to a positive result of B .

5.3.2 The χ^2 test for comparison between two dependent sample proportions

Paired design is often chosen to compare two different treatments in order to control random error due to the variation among individuals. For instance, every two similar subjects are paired, of which one is randomly selected to receive treatment A and another to receive treatment B ; the reaction to the treatment is a binary variable (positive or negative). In general, the results are usually listed with the format of Table 5.7a. And accordingly we have Table 5.7b.

The purpose of this kind of studies is to compare if the probabilities of positive reaction to the two treatments are equal or not, that is, to test

$$H_0: \pi_{c1} = \pi_{r1} \quad H_1: \pi_{c1} \neq \pi_{r1}.$$

Since

$$\pi_{c1} = \pi_{11} + \pi_{21}, \quad \pi_{r1} = \pi_{11} + \pi_{12}$$

the test becomes

$$H_0: \pi_{12} = \pi_{21}, \quad H_1: \pi_{12} \neq \pi_{21}.$$

Table 5.7a The data format for comparison between two treatments under paired design.

Treatment A	Treatment B		Total
	+	−	
+	f_{11}	f_{12}	n_{r1}
−	f_{21}	f_{22}	n_{r2}
Total	n_{c1}	n_{c2}	n (fixed)

Table 5.7b The probability expression of data in Table 5.7a.

Treatment A	Treatment B		Total
	+	-	
+	π_{11}	π_{12}	π_{r1}
-	π_{21}	π_{22}	π_{22}
Total	π_{c1}	π_{c2}	1.0

The easiest way to perform such a test is still the goodness-of-fit χ^2 test. The test statistic to be used is

$$\chi_P^2 = \frac{(f_{12} - e_{12})^2}{e_{12}} + \frac{(f_{21} - e_{21})^2}{e_{21}}, \quad \nu = 2 - 1 = 1. \quad (5.13a)$$

When H_0 is true,

$$\pi_{12} = \pi_{21} \approx \left(\frac{f_{12} + f_{21}}{2} \right) / n,$$

$$e_{12} = n\pi_{12} \approx \frac{f_{12} + f_{21}}{2}, \quad e_{21} = n\pi_{21} \approx \frac{f_{12} + f_{21}}{2}.$$

Substituting the expected frequencies into Eq. (5.13) and then simplify it,

$$\chi_P^2 = \frac{(f_{12} - f_{21})^2}{f_{12} + f_{21}}, \quad \nu = 2 - 1 = 1. \quad (5.13b)$$

When $f_{12} + f_{21}$ is not big enough, the formula of correction for continuity is

$$\chi_P^2 = \frac{(|f_{12} - f_{21}| - 1)^2}{f_{12} + f_{21}}, \quad \nu = 2 - 1 = 1. \quad (5.13c)$$

This test is called McNemer's test. The two figures f_{11} and f_{22} are not utilized because these two cells do not provide any information on the difference between the two treatments when n is fixed; the inference is conditioned on the observed frequencies f_{12} and f_{21} so that it is subject to a kind of conditional method. There are some non-conditional methods which also utilize the information of f_{11} and f_{22} but the calculation is relatively complicated and beyond the scope of this book.

Example 5.5 (Cont'd from Example 5.4) Comparing the two methods, whether the positive rates of test *A* and test *B* are equal in the population?

Solution The equation of (5.13c) is used to calculate the value of χ^2 .

$$\chi^2 = \frac{(|8 - 12| - 1)^2}{8 + 12} = 0.45.$$

Since $\chi^2 = 0.45$, $\chi_{0.05}^2 = 3.84$, $P > 0.05$, H_0 cannot be rejected. We may consider that the positive rates of the two methods are equal.

Both Secs. 5.2 and 5.3.2 are concerning with the comparison between two sample proportions. What is the difference between them? In Sec. 5.2, two independent samples based on a completely random design are compared; while in Sec. 5.3.2 there is only one sample based on a randomized paired design, of which two dependent sub-samples are compared.

Both Secs. 5.3.1 and 5.3.2 are based on paired design for binary variables, and the similar format of data (Tables 5.5a and 5.7a) are shared. In particular, Examples 5.4 and 5.5 have been worked on the same data set. What is the difference between the two? Paragraph 5.3.1 is concerning about the independence between two binary variables, while paragraph 5.3.2 is concerning about the difference between two proportions. Sometimes (not all the times) both tests may be done respectively based on the same data set for different purposes, but obviously, the results have different meaning.

Incorporating Example 5.4 with Example 5.5, one may conclude that the results of the two methods were associated in certain degree; although their results might not always be consistent, their total positive rates were not significantly different.

5.4 The χ^2 Test for $R \times C$ Contingency Table

In Secs. 5.1 and 5.3, the χ^2 tests for data of 2×2 contingency table were discussed. In practice, it is frequently faced that the numbers of rows and/or columns are more than 2 and this kind of tables is called $R \times C$ contingency table. R is the number of rows and C is the number of columns. The 2×2 table is a special case of $R \times C$ table. The χ^2 test for $R \times C$ table includes the χ^2 test for the whole $R \times C$ table and that for split tables.

5.4.1 The χ^2 test for the whole $R \times C$ table

As showed before, the formula (5.7a) can be used for comparison between two independent sample proportions and for testing independence between two binary variables. A straightforward extension of (5.7a) for $R \times C$ contingency table is

$$\chi_P^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{(f_{ij} - e_{ij})^2}{e_{ij}}, \quad (5.14a)$$

where the meaning of e_{ij} , f_{ij} as well as the formulas for f_{ij} and degrees of freedom are the same as before, but for $i = 1, 2, \dots, R$, $j = 1, 2, \dots, C$. This formula can be used to compare R independent samples for a discrete variable with C categories as well as to test the independence between two categorical variables with R and C categories respectively. An equivalent formula for the convenience of calculation is

$$\chi_P^2 = n \left(\sum_i \sum_j \frac{f_{ij}^2}{n_{ri}n_{cj}} - 1 \right) \quad i = 1, 2, \dots, R, \quad j = 1, 2, \dots, C, \quad (5.14b)$$

where n_{ri} is the sub-total of the i th row, n_{cj} is the sub-total of the j th column, n is the grand total.

A straightforward extension of (5.9) is

$$\chi_L^2 = 2 \sum_{i=1}^R \sum_{j=1}^C f_{ij} \ln \left(\frac{f_{ij}}{e_{ij}} \right) \quad (5.14c)$$

which can be used to calculate χ_L^2 for $R \times C$ contingency table.

Example 5.6 In a project of medical study, three samples were randomly selected from the patients suffering from digestive ulcer, stomach cancer and others (control) respectively, of which the blood type of each individual was tested. The results are listed in Table 5.8. Show if the distributions of blood types are different for different populations.

Solution The test hypotheses are

H_0 : The proportions of blood types in three populations are the same.

H_1 : The proportions are different.

Table 5.8 Blood types of patient suffering from different diseases.

Disease status	Blood type			Total
	A	B	O	
Digestive ulcer	679	134	983	1796
Stomach cancer	416	84	383	883
Control	2625	570	2892	6087
Total	3720	788	4258	8766

By formula (5.14b),

$$\chi_P^2 = 8766 \left[\left(\frac{679^2}{1796 \times 3720} + \cdots + \frac{2892^2}{6087 \times 4258} \right) - 1 \right] = 40.543.$$

By formula (5.13c)

$$\chi_L^2 = 2 \left[679 \ln \left(\frac{679}{762.16} \right) + \cdots + 2892 \ln \left(\frac{2892}{2956.7} \right) \right] = 40.64,$$

where the expected frequencies in above equations are calculated by formula (5.6).

The degrees of freedom is $\nu = (3 - 1)(3 - 1) = 4$ according to formula (5.10), and $\chi_{0.05}^2 = 9.488$ so that $P < 0.05$ and hence H_0 is rejected. The conclusion is that the three diseases might have different distributions of blood type.

5.4.2 The χ^2 tests for split $R \times C$ tables

In Example 5.6, the results of χ^2 test show that people suffering from different diseases might have different distributions of blood type. However, it does not mean the distributions are different between any pair of disease populations. The χ^2 tests may be used to decide which pair of populations has different distributions of blood type by the following steps:

- (1) Calculate the frequency distribution of blood type in each population listed in Table 5.9.

Table 5.9 The blood types of patients suffering from different diseases.

Disease status	Blood types			Total
	A	B	O	
Digestive ulcer	679 (37.80%)	134 (7.46%)	983 (54.73%)	1796 (100.0%)
Stomach cancer	416 (47.11%)	84 (9.51%)	383 (43.37%)	883 (100.0%)
Control	2625 (43.12%)	570 (9.36%)	2892 (47.51%)	6087 (100.0%)

Table 5.10 The first split table from Table 5.8: blood types vs. disease status.

Disease status	Blood types			Total
	A	B	O	
Stomach cancer	416	84	383	883
Control	2625	570	2892	6087
Total	3041	654	3275	6970

Table 5.11 The second split table from Table 5.8: blood types vs. disease status.

Disease status	Blood types			Total
	A	B	O	
Digestive ulcer	679	134	983	1796
Stomach cancer and others	3041	654	3275	6970
Total	3720	788	4258	8766

It seems that the frequency distributions of blood type in the groups of stomach cancer and control are relatively close.

- (2) Split these two groups from Table 5.9 to generate Table 5.10. The value of χ_p^2 for Table 5.10 is 5.636, $\nu = 2$ and $\chi_{0.05}^2 = 5.991$. Since $P > 0.05$, the difference between these two groups are not significant, hence we consider the distributions of blood type of stomach cancer group and control group being the same.
- (3) Combine the two groups together and compared in Table 5.11. The value of χ_p^2 is 34.919, $\nu = 2$. Since $\chi_{0.05}^2 = 5.991$, the difference is significant.

Table 5.12 The results of χ^2 tests.

Table	χ_P^2	χ_L^2	ν
Table 5.10	5.636	5.639	2
Table 5.11	34.919	35.001	2
Total	40.555	40.640	4
Table 5.9	40.543	40.640	4

Conclusion is that distribution of blood type among ulcer patients might be different from other. Similar conclusion can be obtained by χ_L^2 (see Table 5.12).

- (4) Further analysis may be done to see which type of blood is different in the two groups of patients. (Details are omitted)

Attention should be paid to the following points: The purpose of splitting a table is to find the difference so that the proportions in the cells help us to decide the way of splitting the initial table; and only once the observed frequency in each cell is allowed to show up in the split tables.

5.4.3 Measurement of association for $R \times C$ table

Suppose there are two categorical variables A and B ; A has R categories and B has C categories; a random sample is cross classified into an $R \times C$ contingency table. After a χ^2 test for independence between A and B , if the null hypothesis is rejected, again, a Pearson's contingency coefficient r_P can be calculated by formula (5.12) to describe the strength of the association between the two variables quantitatively.

Example 5.7 The treatment effect of a medication had been evaluated directly by 170 randomly selected patients to see whether there was any association between the effect and their age and what the strength of association was (see Table 5.13). The value of χ_P^2 for testing independence was 23.582 and $\nu = (3 - 1)(3 - 1) = 4$. Hypothesis of independence was rejected. By formula (5.12), Pearson contingency coefficient was

$$r_P = \sqrt{\frac{23.582}{170 + 23.582}} = 0.35.$$

Table 5.13 Treatment effect and age of 170 patients.

Age	Effect			Total
	No	Better	Recover	
<18	5	32	20	57
18~	30	38	10	78
50~	15	10	10	35
Total	50	80	40	170

The result showed a weak association between the effect of treatment and patients' age.

In χ^2 test of $R \times C$ table, the expected frequency should be greater than or equal to 5. If the expected frequencies in 20% of cells are less than 5 or any single expected frequency is less than 1, the results of χ^2 test will be biased. One solution is to combine the row or column of the cell with its neighbor row or column to increase the expected frequency in the cell until all the expected frequencies greater than or equal to 5. Of course, the combination of rows or columns should be reasonable to the knowledge of subject matter.

5.5 The χ^2 Test for Confirming a Supposed Distribution

In practice, it is required to test if a sample frequency distribution fits a theoretical distribution. Usually a goodness-of-fit χ^2 test is used.

Example 5.8 There was a break out of bacterial dysentery in a place. In order to explore if there was family-clustering in this epidemic, 288 families with 4 family members were interviewed. The data are listed in Table 5.14.

Solution If there was no family-clustering, the number of cases in family would follow a binomial distribution. A χ^2 test was used to test the goodness-of-fit between the actual frequency distribution and the supposed binomial distribution by the following steps:

(1) Test hypothesis

H_0 : The data follow a binomial distribution $B(\pi, n)$.

H_1 : The data do not follow a binomial distribution.

Table 5.14 Goodness-of-fit test for a binomial distribution.

No. of cases per family X	No. of families f_x	Probability P_x	No. of expected families e_x	Statistic χ^2
0	167	0.4396	126.59	12.90
1	51	0.4011	115.52	36.04
2	50	0.1372	39.53	2.77
3	17	0.0209	6.01	29.25
4	3	0.0012	0.35	
Total	288	1.0000	288.00	80.96

(2) Estimating the parameter π

Number of cases = $0 \times 167 + 1 \times 51 + \dots + 4 \times 3 = 214$

$$\pi \approx \frac{\text{Number of cases}}{\text{Total number of people}} = \frac{214}{4 \times 288} = 0.18576.$$

(3) Calculating the probabilities and expected frequencies According to the theory of binomial distribution,

$$P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}.$$

For example, the probability of the event of no any case in a family is

$$P(X = 0) \approx \binom{4}{0} (0.18576)^0 (1 - 0.18576)^{4-0} = 0.43955.$$

And the expected number of families without any case is

$$e_0 = nP(X = x) \approx 288 \times 0.43955 = 126.59.$$

(4) Calculating χ^2 value and degrees of freedom The part of χ^2 contributed by the cell of $X = 0$ is

$$\chi_0^2 = \frac{(f_0 - e_0)^2}{e_0} = \frac{(167 - 126.59)^2}{126.59} = 12.90.$$

The others can be calculated in the same way except the last two cells.

Since the number of expected families corresponding to $X = 4$ is much less than 5, the cell is combined with that of $X = 3$. Then

$$f_3 = 17 + 3 = 20, \quad e_3 = 6.01 + 0.35 = 6.36,$$

$$\chi_3^2 = \frac{(f_3 - e_3)^2}{e_3} = \frac{(20 - 6.36)^2}{6.36} = 29.25.$$

All results are listed in Table 5.14. Finally the total χ^2 value is 80.99. The number of groups in this example was 4 and a sample incidence rate was used as the estimation of π so that $\nu = 4 - 1 - 1 = 2$.

- (5) **Decision and conclusion** Since $\chi_{0.05(2)}^2 = 5.99$, $P < 0.05$, H_0 is rejected. We might say that the number of patients per family did not follow a binomial distribution; the infection among family members were not independent of each other, there might exist family-clustering.

5.6 Hypothesis Testing for Two Standardized Rates

In Chap. 1, to compare two sets of age-specific rates, the methods of standardization were introduced for adjustment of the crude rates. In Example 1.6, the direct standardized rate was 19.2% at place *A* and 16.3% at place *B*. Suppose the data came from a sampling study and the purpose of the study was to compare the difference of standardized mortality rates, a hypothesis testing was needed to see if the difference was of statistical significance.

5.6.1 Test for direct standardized rates

Example 5.9 (Cont'd of Example 1.6) Test the two direct standardized mortalities at place *A* and *B* to see whether there is statistically significant difference.

Solution The test hypotheses are

H_0 : The two population standardized mortality rates are equal.

H_1 : The two population standardized mortality rates are not equal.

- (1) Calculate the combined estimation of the age-specific mortality rates of the two places.

$$p_i = \frac{d_{Ai} + d_{Bi}}{n_{Ai} + n_{Bi}}, \quad (5.15)$$

where d_{Ai} and n_{Ai} are the numbers of deaths and the population in i th age group of place A ; similarly, d_{Bi} and n_{Bi} are those of place B . $n_{Ai} + n_{Bi} = n_i$, $d_{Ai} + d_{Bi} = d_i$. These are all listed in Table 5.15.

- (2) Calculate the variance of the difference between two age-specific mortality rates, denoted by s_i^2 .

$$s_i^2 = p_i(1 - p_i) \left(\frac{1}{n_{Ai} + n_{Bi}} \right). \quad (5.16)$$

The results are listed in column 9 of Table 5.15.

- (3) Calculate the variance of the difference between two standardized mortality rates, denoted by s^2 .

$$s^2 = \frac{\sum_i h_i^2 s_i^2}{(\sum_i h_i)^2}, \quad (5.17)$$

where h_i is the standard population in i th age group. In Example 1.6, the sum of two populations was used as the standard population so that column 10 is the same as column 4 in Table 5.15. According to the totals of column 11 and 10,

$$s^2 = \frac{\sum_i h_i^2 s_i^2}{(\sum_i h_i)^2} = \frac{1205.67}{(11298)^2} = 9.45 \times 10^{-6}.$$

- (4) Calculate the value of Z . As we have known before, the standardized mortality rates P'_{Ai} and P'_{Bi} are all linear combinations of age specific mortality rates,

$$P'_A = \frac{\sum_i h_i P_{Ai}}{\sum_i h_i}, \quad P'_B = \frac{\sum_i h_i P_{Bi}}{\sum_i h_i}.$$

For large sample, when H_0 is true.

$$Z = \frac{P'_A - P'_B}{\sqrt{s^2}} \sim N(0, 1). \quad (5.18)$$

Table 5.15 Calculation on the test for two direct standardized rates.

Age (year)	Population			Number of deaths			Pooled mortality rate p_i (%)	Variance of difference $s_i^2 \times 10^{-4}$	Standard population h_i	$(10)^2 \times (9) h_i^2 s_i^2$
	Place A n_{Ai}	Place B n_{Bi}	Total n_i	Place A d_{Ai}	Place B d_{Bi}	Total d_i				
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
0~	400	286	686	2	1	3	4.37	0.26	686	12.24
15~	2000	238	2238	10	1	11	4.92	0.23	2238	115.20
30~	2000	794	2794	15	5	20	7.16	0.13	2794	101.48
45~	800	2000	2800	8	18	26	9.29	0.16	2800	125.44
60~	400	2000	2400	16	70	86	35.83	1.04	2400	599.04
75~	80	300	380	12	36	48	126.32	17.47	380	252.27
Total	5680	5618	11298	63	131	194			11298	1205.67

In this example, $P'_A = 19.2\%$ and $P'_B = 16.3\%$

$$Z = \frac{P'_A - P'_B}{\sqrt{s^2}} = \frac{0.0192 - 0.0163}{\sqrt{9.45 \times 10^{-6}}} = 0.94.$$

(5) Decision and conclusion. Since $Z < 1.96$, $P > 0.05$, H_0 is not rejected and we might say that the standardized mortality rates in two places are different.

5.6.2 Test for SMR from indirect standardization

Example 5.10 (Cont'd of Example 1.6) Test the two indirect standardized mortalities at places A and B to check whether there is statistically significant difference.

Solution A goodness-of-fit χ^2 test will be applied for places A and B separately. For place A , the test hypotheses are

H_0 : The age-specific mortality rates of place A are the same as those of the standard population,

H_1 : The age-specific mortality rates of place A are different from those of standard population.

The actual number of deaths at place A is regarded as the observed frequency f_{A1} and the expected number of deaths estimated by the indirect method as the expected frequency e_{A1} ; the actual number of survivors is regarded as the observed frequency $f_{A2} = n_A - f_{A1}$, and the corresponding expected frequency $e_{A2} = n_A - e_{A1}$. When H_0 is true, the statistic

$$\chi_A^2 = \sum_{i=1}^2 \frac{(|f_{Ai} - e_{Ai}| - 0.5)^2}{e_{Ai}} \sim \chi^2 \text{ distribution.} \quad (5.19a)$$

Since the standard population is given, rather than estimated by sample data, the degrees of freedom $\nu = 2 - 1 - 0 = 1$.

In Example 1.6, by (5.19a), $f_{A1} = 63$, $e_{A1} = 58.12$; $f_{A2} = 5680 - 63 = 5617$, $e_{A2} = 5680 - 58.12 = 5621.88$,

$$\chi_A^2 = \sum_{i=1}^2 \frac{(|f_{Ai} - e_{Ai}| - 0.5)^2}{e_{Ai}}$$

$$\begin{aligned}
 &= \frac{(|63 - 58.12| - 0.5)^2}{58.12} + \frac{(|5617 - 5621.88| - 0.5)^2}{5621.88} \\
 &= \frac{4.38^2}{58.12} + \frac{4.38^2}{5621.88} \approx \frac{4.38^2}{58.12} = 0.33.
 \end{aligned}$$

From the above calculation, one can see that the second term of the statistic can always be omitted because the numerator e_{A2} is always very large. Therefore, (5.19a) can be simplified as a single term as follows:

$$\chi_A^2 = \frac{(|f_A - e_A| - 0.5)^2}{e_A} \sim \chi^2 \text{ distribution}, \quad (5.19b)$$

where f_A and e_A refer to the observed frequency and expected frequency of deaths respectively.

In Example 1.6, the χ^2 value for the test is less than $\chi_{0.05(1)}^2 = 3.84$, so that H_0 is not rejected.

Similarly, we can have the test for place B . Accordingly,

H_0 : The age-specific mortality rates of place B are the same as those of the standard population,

H_1 : The age-specific mortality rates of place B are different from those of standard population.

By (5.19b), $f_B = 131$, $e_B = 142.3$,

$$\chi_B^2 = \frac{(|f_B - e_B| - 0.5)^2}{e_B} = \frac{(|131 - 142.3| - 0.5)^2}{142.3} = 0.82, \quad \nu = 1.$$

The χ^2 value is also less than $\chi_{0.05(1)}^2 = 3.84$ so that H_0 is not rejected either.

As a summary, the mortality rates at both places A and B are not different from the standard population, and hence the two standardized mortality rates can be regarded as equal to each other.

5.7 Fisher's Exact Test for 2×2 Table

The χ^2 distribution is a continuous distribution. The statistic χ^2 defined by Eq. (5.7a) only approximate to a χ^2 variable when the null hypothesis is true. The approximation is not strictly valid when some of the expected frequencies are small, such as smaller than 1. By experience, when the

Table 5.16 The results of treatment to thromboangiitis angiitis patients.

Groups	Recovery	No recovery	Total
New treatment	6(<i>a</i>)	1(<i>b</i>)	7(<i>n_{r1}</i>)
Control	1(<i>c</i>)	4(<i>d</i>)	5(<i>n_{r2}</i>)
Total	7(<i>n_{c1}</i>)	5(<i>n_{c2}</i>)	12(<i>n</i>)

total sample size is smaller than 40, the bias cannot be simply adjusted. Therefore, in this situation, we may directly estimate the exact probability for the occurrence of each possible event conditioning on the given marginal frequencies, by the theory of hypergeometric distribution. This method is so-called Fisher's exact test being suggested by R. A. Fisher (1934).

Now we would introduce the Fisher's exact test through the following example.

Example 5.11 12 patients suffering from thromboangiitis angiitis were randomly divided into two groups, receiving a new treatment and the routine treatment (control) respectively. The results is listed in Table 5.16. The question is whether the recovery rates of the two groups are significantly different.

5.7.1 Setting up the testing hypotheses

Back to the data with the format of Table 5.1a and their probability expression in Table 5.1b, the hypotheses we want to test are still

$$H_0: \pi_1 = \pi_2 \text{ (Two population probabilities are equal)}$$

$$H_1: \pi_1 \neq \pi_2 \text{ (Two population probabilities are not equal).}$$

5.7.2 The conditional probabilities

The basic idea here is to have the inference conditioning on the given marginal sub-totals $n_{r1}, n_{r2}, n_{c1}, n_{c2}$. We will introduce the procedures to list all the possible events and calculate their probabilities.

- (1) List all the possible 2×2 tables given the values of $n_{r1}, n_{r2}, n_{c1}, n_{c2}$.

For Example 5.11, all the possible 2×2 tables are listed as follows under the condition of $n_{r1} = 7, n_{r2} = 5, n_{c1} = 7, n_{c2} = 5$. One can see

that there is only one 2×2 table corresponding to each value of a , in other words, once the value of a is fixed, the values of b , c and d must be fixed.

(1) $a = 2$	(2) $a = 3$	(3) $a = 4$	(4) $a = 5$	(5) $a = 6$	(6) $a = 7$																								
<table><tr><td>2</td><td>5</td></tr><tr><td>5</td><td>0</td></tr></table>	2	5	5	0	<table><tr><td>3</td><td>4</td></tr><tr><td>4</td><td>1</td></tr></table>	3	4	4	1	<table><tr><td>4</td><td>3</td></tr><tr><td>3</td><td>2</td></tr></table>	4	3	3	2	<table><tr><td>5</td><td>2</td></tr><tr><td>2</td><td>3</td></tr></table>	5	2	2	3	<table><tr><td>6</td><td>1</td></tr><tr><td>1</td><td>4</td></tr></table>	6	1	1	4	<table><tr><td>7</td><td>0</td></tr><tr><td>0</td><td>5</td></tr></table>	7	0	0	5
2	5																												
5	0																												
3	4																												
4	1																												
4	3																												
3	2																												
5	2																												
2	3																												
6	1																												
1	4																												
7	0																												
0	5																												

- (2) Calculate the conditional probability $P(f_{11} = a | n_{r1}, n_{r2}, n_{c1}, n_{c2})$ when H_0 is true.

On the one hand, because H_0 means that there is no difference between the two groups, among the recovery patients (n_{c1}), any one has the same chance to be allocated into the new treatment group or the control group. The number of possible ways of allocating $f_{11} = a$ patients into the new group is the number of possible combinations, $\binom{n_{c1}}{a}$. Similarly, among the un-recovery patients (n_{c2}), the number of possible ways of allocating $f_{12} = b$ patients into the new treatment group is the number of possible combinations, $\binom{n_{c2}}{b}$.

On the other hand, by randomization, among all the patients (n), the number of possible ways of allocating n_{r1} patients into the new treatment group is the number of possible combinations, $\binom{n}{n_{r1}}$. Therefore, the probability of having $f_{11} = a$ recovery patients in the new treatment group (n_{r1}) is

$$P(a | n_{r1}, n_{r2}, n_{c1}, n_{c2}) = \frac{\binom{n_{c1}}{a} \binom{n_{c2}}{b}}{\binom{n}{n_{r1}}} = \frac{n_{r1}! n_{r2}! n_{c1}! n_{c2}!}{a! b! c! d! n!} \quad (5.20)$$

where “!” means “factorial”,

$$n! = n \times (n - 1) \times \cdots \times 2 \times 1.$$

For Example 5.11,

$$\begin{aligned} P(a = 6 | n_{r1} = 7, n_{r2} = 5, n_{c1} = 7, n_{c2} = 5) \\ = \frac{7! 5! 7! 5!}{6! 1! 1! 4! 12!} = 0.044192. \end{aligned}$$

Table 5.17 Probabilities of all the possible events.

$P(a = 2)$	$P(a = 3)$	$P(a = 4)$	$P(a = 5)$	$P(a = 6)$	$P(a = 7)$
0.02651515	0.22095959	0.44191919	0.28515152	0.04419192	0.00126263

This is just a probability of one possible event. In the same way, the conditional probabilities of all possible 2×2 tables have been calculated and listed in Table 5.17, where the conditions of $n_{r1} = 6, n_{r2} = 1, n_{c1} = 1, n_{c2} = 5$ have been omitted for short.

5.7.3 *P value and conclusion*

According to the definition of P value in general hypothesis testing, it is the probability of all possible events, including the current event as well as the events that are more extremely departure from the null hypothesis.

For Example 5.11, if it is a one-side test, corresponding to the alternative hypothesis $H_1: \pi_1 > \pi_2$, the more extreme situation than $a = 6$ is $a = 7$ so that the P value is

$$\begin{aligned} P(a \geq 6 | n_{r1} = 7, n_{r2} = 5, n_{c1} = 7, n_{c2} = 5) \\ = 0.044192 + 0.00126263 = 0.0454545. \end{aligned}$$

Since $P < 0.05$, the null hypothesis $H_0: \pi_1 = \pi_2$ is rejected and the recovery rate of the new medication is higher than that of the control.

However, if the purpose of study is only to check if the recovery rates of the two treatments are different, i.e. $H_1: \pi_1 \neq \pi_2$, then the “extreme events” should include $a = 7$ as well as $a = 2$, because the event of $a = 2$ (recovery rates 2/7 versus 5/5) is also more extremely departure from $H_0: \pi_1 = \pi_2$ than $a = 6$ (recovery rates 6/7 versus 1/5) is. Therefore, the P value of two-side test ought to be

$$P(a = 6) + P(a = 2) + P(a = 7) = 0.071977$$

which is greater than 0.05 so that $H_0: \pi_1 = \pi_2$ is not rejected. This means that if we do not have enough evidence to support a one-side test, then according to the data alone the difference of recovery rates in the two groups is not of statistical significance.

Why we choose “ $a = 7$ ” and “ $a = 2$ ” to calculate the probability (P value)? In this example, events more extremely are equivalent to increase

the current value of statistic χ^2 . That is to say, the difference between expect frequencies and theoretical frequencies become larger. When “ $a = 7$ ” or “ $a = 2$ ”, from the special formula,

$$\chi^2 = \frac{(ad - bc)^2 n}{(a + b)(c + d)(a + c)(b + d)}$$

we know that, χ^2 value will be larger than that when “ $a = 6$ ”, so we could use “ $a = 7$ ” and “ $a = 2$ ” to calculate the P value.

5.8 Computerized Experiments

Experiment 5.1 The χ^2 test for comparing independent samples and testing independence between two variables Example 5.6 is used as an example. The SAS program is listed in Program 5.1.

In Program 5.1, Line 01 sets 300 lines print out per page. Lines 02–08 construct an SAS data set named AAA. Data include two categorical variables (A and B) and one frequency variable. Lines 09–13 use the procedure FREQUENCY in SAS to do χ^2 test. Line 12 identifies COUNT as frequency variable.

Experiment 5.2 The goodness-of-fit χ^2 test for binomial distribution Example 5.8 is used as an example. SAS program is given in Program 5.2.

In Program 5.2, lines 01–11 construct an SAS data set and print it out. Lines 12–24 estimate the parameter π in binomial distribution. Lines 25–29 calculate the probabilities for different values of X . Lines 30–38 calculate combination of $X = 3$ and $X = 4$. Lines 39–42 calculate the expected

Program 5.1 Comparison of two independent sample proportions and test for independence.

Line	Program	Line	Program
01	OPTIONS PS=300;	08	;
02	DATA AAA;	09	PROC FREQ DATA=AAA;
03	INPUT A B COUNT @@;	10	TABLES A*B/CHISQ NOCOL
04	CARDS;	11	CELLCHI2 NOPERCENT NOCUM;
05	1 1 679 1 2 134 1 3 983	12	WEIGHT COUNT;
06	2 1 416 2 2 84 2 3 383	13	RUN;
07	3 1 2625 3 2 570 3 3 2892		

Program 5.2 The Goodness-of-fit χ^2 test for a binomial distribution.

Line	Program	Line	Program
01	DATA A;	25	DATA D E;
02	DO X=0 TO 4;	26	SET C;
03	INPUT FX @@ ;	27	IF X=0 THEN PX =PROBBNML(P,4,0);
04	T=X*FX;	28	ELSE PX=PROBBNML(P,4,X)- PROBBNML(P,4,(X-1));
05	TOT=4*FX;	29	IF X < 3 THEN OUTPUT D; ELSE OUTPUT E;
06	OUTPUT;	30	PROC MEANS DATA=E SUM NOPRINT;
07	END;	31	VAR FX PX;
08	CARDS;	32	OUTPUT OUT=F SUM =SUMFX SUMPX;
09	167 51 50 17 3	33	DATA G;
10	;	34	KEEP X FX PX;
11	PROC PRINT; VAR X FX;	35	SET F;
12	PROC MEANS SUM NOPRINT;	36	X=3.4;
13	VAR T TOT;	37	FX=SUMFX;
14	OUTPUT OUT=B SUM=SUMT SUMTOT;	38	PX=SUMPX;
15	DATA P;	39	DATA H;
16	KEEP X P;	40	SET D G;
17	SET B;	41	EX=288*PX;
18	DO X=0 TO 4;	42	CHISQ=(FX-EX)**2/EX;
19	P=SUMT/SUMTOT;	43	PROC TABULATE DATA=h ORDER=DATA;
20	OUTPUT;	44	CLASS X;
21	END;	45	VAR FX PX EX CHISQ;
22	DATA C ;	46	TABLE X ALL, FX PX EX CHISQ;
23	MERGE A P;	47	RUN;
24	BY X;		

frequencies for different values of X and χ^2 values. Lines 43–47 print the final results in a table format. The results are listed in Table 5.18.

Experiment 5.3 The goodness-of-fit χ^2 test for Poisson distribution The red blood cell count (RBC) in each view of microscope is recorded and the data are listed in Table 5.19. SAS program is given in Program 5.3.

In Program 5.3, lines 01–10 construct an SAS data A. Lines 11–22 calculate the parameter λ of Poisson distribution. Lines 23–33 calculate the

Table 5.18 The results of Program 5.2.

X	FX	PX	EX	CHISQ
0	167.00	0.44	126.59	12.90
1	51.00	0.40	115.52	36.04
2	50.00	0.14	39.53	2.77
3.4	20.00	0.02	6.36	29.29
ALL	288.00	1.00	288.00	81.00

Table 5.19 RBC distribution.

RBC	0	1	2	3	4	5	6	7	8	9
Views	11	36	76	80	74	58	38	17	6	4

Program 5.3 The Goodness-of-fit χ^2 test for a poisson distribution.

Line	Program	Line	Program
01	DATA A;	21	KEEP X F SUMF LAMDA;
02	DO X=0 TO 9;	22	MERGE A C; BY X;
03	INPUT F @@;	23	DATA E;
04	T=X*F;	24	SET D;
05	OUTPUT;	25	IF X=0 THEN P=POISSON (LAMDA,0);
06	END;	26	ELSE IF X<9 THEN P=POISSON(LAMDA,X) -POISSON(LAMDA,(X-1));
07	CARDS;	27	ELSE P=1-POISSON (LAMDA,(X-1));
08	11 36 76 80 74 58 38 17 6 4	28	RETAIN SCHISQ 0 CP 0 ;
09	;	29	CP=CP+P;
10	PROC PRINT ; VAR X F;	30	T=SUMF*P;
11	PROC MEANS SUM NOPRINT;	31	CHISQ=((F-T)**2)/T;
12	VAR T F;	32	SCHISQ=SCHISQ+CHISQ;
13	OUTPUT OUT=B SUM=SUMT SUMF;	33	OUTPUT;
14	DATA C;	34	PROC PRINT;
15	KEEP X SUMF LAMDA;	35	VAR X F P T CHISQ ;
16	SET B;	36	PROC MEANS SUM;
17	DO X=0 TO 9;	37	VAR CHISQ;
18	LAMDA=SUMT/SUMF;	38	OUTPUT OUT=B SUM=SUMCHI;
19	OUTPUT; END;	39	RUN;
20	DATA D;		

Table 5.20 Results of Program 5.3.

X	F	P	T	CHISQ
0	11	0.02698	10.7937	0.00394
1	36	0.09748	38.9923	0.22963
2	76	0.17607	70.4299	0.44053
3	80	0.21202	84.8093	0.27272
4	74	0.19148	76.5934	0.08781
5	58	0.13835	55.3387	0.12798
6	38	0.08330	33.3185	0.65778
7	17	0.04299	17.1947	0.00221
8	6	0.01941	7.7645	0.40099
9	4	0.01191	4.7649	0.12278
Total				2.34637

probabilities, expected frequencies and χ^2 values for different values of X . Lines 34–39 print the results of the test (Table 5.20).

Experiment 5.4 Two types of error in χ^2 test Assume a binomial population with $\pi_1 = 0.3$ and another binomial population with $\pi_2 = 0.4$. 2000 samples are randomly generated from the first population with a sample size $n = 60$. The χ^2 values of 2000 samples are calculated and summarized by a histogram; the conclusion of hypothesis testing for each sample will be made at a significant level $\alpha = 0.05$. The SAS program is given in Program 5.4.

In Program 5.4, lines 01–09 generate 2000 samples randomly from a binomial population ($\pi_1 = 0.3, n = 60$), and at the same time calculate the χ^2 values under two binomial distributions $B(0.3, 60)$ and $B(0.4, 60)$ respectively. Lines 10 and 11 show a histogram of χ^2 values. Lines 12–15 describe the χ^2 values by procedure UNIVARIATE. Lines 16–25 give scores for type I error and type II error at a significant level of $\alpha = 0.05$. Lines 26–28 estimate the probabilities of two types of error by procedure MEANS. Finally, lines 29–37 calculate and print the results.

5.9 Practice and Experiments

1. In the goodness-of-fit test, why does it only consider the upper tail area of chi-square distribution?

Program 5.4 Sampling experiments of two types of error in χ^2 test.

Line	Program	Line	Program
01	DATA A;	19	IF CHI1 >= 3.84 THEN
02	DO I=1 TO 2000;	20	CONUT1=1;
03	X=RANBIN(0,60,0.30) ;	21	ELSE COUNT1=0;
04	P=X/60;	22	S1=S1+COUNT1;
05	Q=1-P;	23	IF CHI2 < 3.84 THEN
06	CHI1=(60*P-18)**2/18	24	COUNT2=1;
	+ (60*Q-42)**2/42;	25	ELSE COUNT2=0;
07	CHI2=(60*P-24)**2/24	26	S2=S2+COUNT2;
	+ (60*Q-36)**2/36;	27	PROC MEANS SUM
08	OUTPUT;	28	NOPRINT;
09	END;	29	VAR COUNT1 COUNT2;
		30	OUTPUT OUT=D
10	PROC CHART;	31	SUM=TOTAL1 TOTAL2;
11	HBAR CHI1 CHI2/TYPE=FREQ;	32	DATA E;
12	PROC UNIVARIATE DATA=A	33	SET D;
	PLOT FREQ NOPRINT;	34	N=60;
13	OUTPUT OUT=B P95=P951 P952;	35	ALPHA1=TOTAL1/2000;
14	VAR CHI1 CHI2;	36	ALPHA=0.05;
15	DATA C;		BETA=TOTAL2/2000;
16	SET A;		POWER=1-BETA;
17	S1=0;		PROC PRINT;
18	S2=0;		RUN;

2. A health care doctor in a factory plans to study the incidence probabilities of workers in five workshops. Can the chi-square test be used to compare the difference among these five incidence probabilities? Give your reasons.
3. There are 6 faces of a dice and each face is marked with a figure of 1, 2, 3, 4, 5 and 6. Toss 60 times of this dice repeatedly in an experiment. The results are listed in the following table. How to evaluate if this dice is even or not?

Figure	1	2	3	4	5	6
Frequency	8	8	5	10	14	15

Table 5.21 The frequency distribution of industrial accidents in 10 years period.

Accidents per month	Frequency of months	Accidents per month	Frequency of months
0	2	6	10
1	10	7	6
2	15	8	2
3	30	9	1
4	28	≥ 10	1
5	15		

Table 5.22 The frequency distribution of numbers of bacteria in 100 samples.

No. of bacteria per sample	Frequency of samples	No. of bacteria per sample	Frequency of samples
0	15	4	5
1	30	5	4
2	25	6	1
3	20	7	0

Table 5.23 The positive rates of a test for two groups of children.

Group	Total	Positive	Positive rate (%)
Urban	58	18	31.0
Rural	147	26	17.7

- The numbers of industrial accidents were recorded monthly in a hospital in the past 10 years (Table 5.21). Does the monthly frequency follow a Poisson distribution?
- A sampling inspection of the products in a food factory reported the results of 100 samples in Table 5.22. Does the number of bacteria per sample follow a Poisson distribution?
- A test was used for examination of two groups of children randomly selected from urban and rural areas of a city respectively. The results are listed in Table 5.23. Are the positive rates of two groups of children significantly different?

Table 5.24 The eye vision in different age groups of population.

Vision	Age group (year)				Total
	5-10	11-20	21-40	41-	
≤ 0.6	4	9	39	147	199
0.7-0.9	11	37	22	94	164
1.0-1.2	143	317	182	139	781
≥ 1.5	411	1183	355	160	2109
Total	569	1546	598	540	3253

Table 5.25 The results of a paired case-control study of MI and smoking.

Control	MI		Total
	Smoking	Non-smoking	
Smoking	88	19	107
Non-smoking	69	24	93
Total	157	43	200

7. Table 5.24 lists the distribution of age and vision. Try to analyze whether there is a relationship between age and vision and calculate the contingency coefficient.
8. Suppose that $\{X_1, X_2, \dots, X_n\}$ is an independent random sample from a normal distribution $N(\mu, \sigma^2)$. Let the sample mean $\bar{X} = \sum X_i/n$ and sample variance $S^2 = \sum (X_i - \bar{X})^2/(n-1)$. Try to prove that $(n-1)S^2/\sigma^2$ has a chi-square distribution with $(n-1)$ degrees of freedom.

$$(\text{Hint: } (n-1)S^2 = \sum [(X_i - \mu) - (\bar{X} - \mu)]^2 = \sum (X_i - \mu)^2 - n(\bar{X} - \mu)^2.)$$

9. Table 5.25 is the data from a paired case-control study of the relationship between myocardial infarction (MI) and smoking. The frequencies in the table are the numbers of pairs. The questions are if MI patients have a higher probability of smoking than controls do and if there is an association between MI and smoking.

10. In order to evaluate the effect of a sensitivity enhancer of radiation, 21 nude mice with tumor were randomly assigned to two groups. Ten mice in the first group received radiotherapy only and other 11 mice in the second group received radiotherapy with sensitivity enhancer. After a period of treatment, three in the first group and six in second group were effective. Try to answer if the effect of sensitivity enhancer is of statistical significance?

(1st edn. Qing Liu, Jiqian Fang; 2nd edn. Yuantao Hao, Te Deng, Yong Huang, Jiqian Fang)

Chapter 6

Further Discussion on Hypothesis Test

As we know, if a hypothesis test miss concludes with “the difference is statistically significant” when H_0 is true, then we say a type I error is made and the probability of type I error is less than or equal to α , the level of the test; if a hypothesis test miss concludes with “the difference is not statistically significant” when H_1 is true, then we say a type II error is made and the probability of type II error is denoted with β , and accordingly, the probability of correctly recognizing the difference, $1 - \beta$, is called the power of the test. It is not unusual in clinical trials to conclude with “the difference is not statistically significant” so that it is necessary to evaluate the power of the test.

The power of a test is closely related with the sample size so that the researchers need to take care of the sample size at the stage of design. The sample size depends on the purpose of the study and the characteristics of the problem, rather than on “convenience”. For instance, in clinical trials a common tendency is the sample size being too small such that the superiority of some new medications can hardly be revealed. To estimate the sample size in advance is really important.

The estimation of sample size at the design stage and the evaluation of power after a test are closely linked to each other. The related concepts and algorithm will be introduced in this chapter.

With the development of modern society, the traditional methods of hypothesis testing and statistical inference are not enough to meet the needs of clinical trials. The commonly used newly developed statistical inference methods for the clinical trials will be introduced in this chapter as well, including the non-inferiority test, equivalence test and permutation test.

6.1 Two Types of Error and Power

6.1.1 The probabilities of two types of error

In fact, the hypothesis test is to decide between the null hypothesis H_0 and the alternative hypothesis H_1 , to which sometimes mistake may inevitably happen.

For example, to test the effectiveness of a new medication for high blood fat, n patients were randomly selected among volunteers with similar conditions, of whom the values of decrement in blood fat were measured respectively. Suppose the average decrement of blood fat with routine medication was μ_0 . To check if the new medication was better, one turned to the hypothesis test

$$H_0 : \mu = \mu_0, \quad H_1 : \mu > \mu_0. \quad (6.1)$$

Assume the decrement in blood fat X follows a normal distribution $N(\mu, \sigma^2)$. When σ is known, the statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \quad (6.2)$$

can be used. When H_0 is true, $Z \sim N(0, 1)$. Given the level of the test α , denote the one-sided critical value Z_α of the standard normal distribution corresponding to the area of the upper tail α . Substitute the sample mean \bar{X} , σ and n into Eq. (6.2) to calculate the value of Z . The decision rule is: When $Z \geq Z_\alpha$, reject H_0 , otherwise, not reject H_0 .

One may have two types of error: type I error is to claim the new medication better than the routine one when both effects are equal. That is, to reject H_0 when H_0 is true (to reject a true H_0); type II error is to claim the new medication equivalent to the routine one when it is really better, that is, not to reject H_0 when H_1 is true (not to reject a false H_0).

Usually the possibilities of these two types of error are measured with two probabilities:

$$\text{Probability of type I error} = P(\text{reject } H_0 | H_0 \text{ is true}) \leq \alpha, \quad (6.3)$$

$$\text{Probability of type II error} = P(\text{not reject } H_0 | H_1 \text{ is true}) \leq \beta, \quad (6.4a)$$

or

$$\text{Probability of type II error} = P(\text{not reject } H_0 | H_0 \text{ is false}) \leq \beta. \quad (6.4b)$$

If the diagnosis of certain disease is also regarded as a problem of hypothesis test,

H_0 : The condition is not different from normal (without disease),

H_1 : The condition is different from normal (with disease),

then the false positive rate is controlled to be below α , and the false negative rate is controlled to be below β .

In hypothesis test, the value of α is given in advance, say $\alpha = 0.05$ or 0.01 , sometimes $\alpha = 0.10$. However, it is not allowed to give the value of α at one's own will. Instead, it should be determined according to the imperilment of type I error. Back to the example of clinical trials, type I error may lead us to inappropriately regard the new medication as innovation and give up the routine one and put into production for nothing. It is impossible to make $\alpha = 0$. The acceptable size of α should be determined incorporating the background of the problem.

In theory, the value of β also needs to be determined in advance. Unfortunately, it is often ignored in practice. In fact, at the design stage, without the value of β the sample size can hardly be estimated. The acceptable size of β should be determined according to the imperilment of type II error. For example, in clinical trial, type II error may lead us to neglect a better medication so that it cannot be produced and utilized. It is impossible to make $\beta = 0$. The acceptable size of β should also be determined incorporating the background of the problem.

6.1.2 Power

To a hypothesis test, if the probability of type II error is controlled to be below β , then the power of this test in finding out the difference will be above $1 - \beta$, that is,

$$\text{The power of the test} = P(\text{reject } H_0 | H_1 \text{ is true}) \geq 1 - \beta. \quad (6.5)$$

In clinical trials, the power is the probability of a good medication being recognized; in the problem of diagnosis, the power is the probability of a disease being diagnosed.

6.2 The Four Elements Affecting the Power

The power of a hypothesis test is affected by at least four elements: the real difference, the variation among individuals, sample size and the test level α .

For convenience of illustration, let us still take the clinical trial of a new medication for high blood fat as an example.

From (6.2), $Z \geq Z_\alpha$ is equivalent to

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq Z_\alpha \quad (6.6)$$

or

$$\bar{X} \geq \mu_0 + Z_\alpha \frac{\sigma}{\sqrt{n}}. \quad (6.7)$$

In other words, the decision can be changed as follows: if (6.7) holds, then reject H_0 ; otherwise, not reject H_0 .

Now let us discuss qualitatively the effect of the four elements to the power through the distributions of sample mean \bar{X} .

6.2.1 Larger difference leading to larger power

If X follows a normal distribution $N(\mu, \sigma^2)$, then the sample mean

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right). \quad (6.8)$$

When H_0 is true,

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right). \quad (6.9)$$

When H_1 is true,

$$\bar{X} \sim N\left(\mu_0 + \delta, \frac{\sigma^2}{n}\right), \quad (6.10)$$

where δ is the real difference between μ and μ_0 .

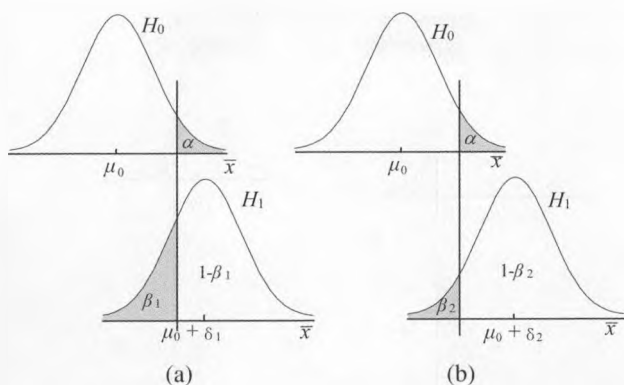


Fig. 6.1 Larger difference leading to larger power. (a) small δ_1 , (b) larger δ_2 .

The upper plots of Fig. 6.1 are the density functions of \bar{X} when H_0 is true, the lower plots are the density functions of \bar{X} when H_1 is true; the vertical lines at $\bar{X} = \mu_0 + Z_\alpha \alpha / \sqrt{n}$ show the dividing points. According to the decision rule, the areas of the right tails with shadow in the upper plots are all equal to α , the probability of type I error; and the areas of the right “tails” without shadow in the lower plots are the powers. The difference between (a) and (b) of Fig. 6.1 is the size of δ , $\delta_1 < \delta_2$ and $1 - \beta_1 < 1 - \beta_2$.

One can see, when the other elements keep the same, the larger existing difference leads to a larger power, that is, more chance to detect the difference.

6.2.2 Smaller variation or larger sample size leading to larger power

In Fig. 6.2, the standard deviation of \bar{X} in (a) is larger than that in (b) such that the density function in (a) looks shorter and fatter than that in (b), and the dividing point in (a) is farther from μ_0 than that in (b). One can see, when the other elements keep the same, the smaller standard deviation makes a less overlap between the two density functions of \bar{X} under H_0 and H_1 .

From (6.8), when the standard deviation is smaller or the sample size is larger, the standard deviation of sample mean will be smaller, and hence the power will be larger.

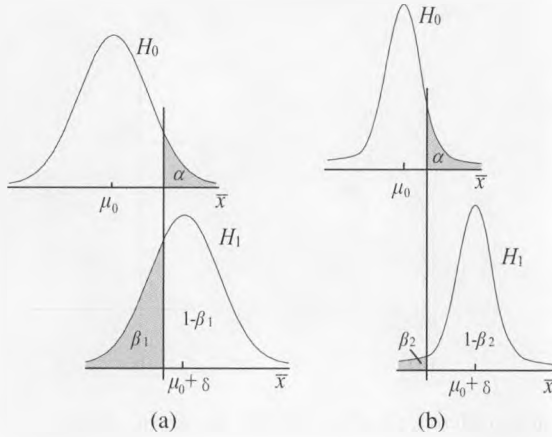


Fig. 6.2 Smaller variation or larger sample size leading to larger power. (a) larger variation or smaller sample size, (b) Smaller variation or larger sample size.

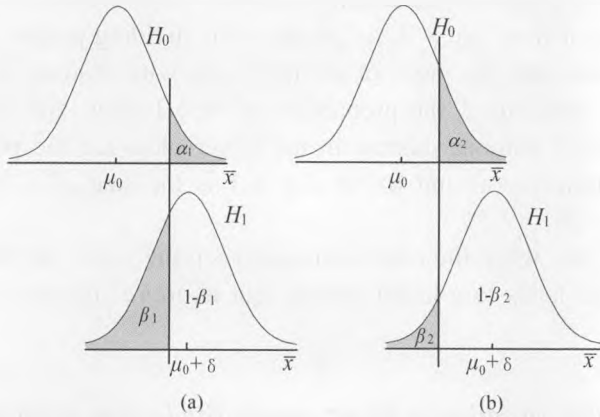


Fig. 6.3 Larger α leading to larger power. (a) smaller α_1 , (b) larger α_2 .

6.2.3 Larger α leading to larger power

From (6.7), the distance between the dividing point and μ_0 is proportional to Z_α ; when α is larger, Z_α will be smaller and the dividing point will be close to μ_0 . The differences between (a) and (b) in Fig. 6.3 are $\alpha_1 < \alpha_2$ and $1 - \beta_1 < 1 - \beta_2$. One can see, when the other elements keep the same, if the limitation on the probability of type I error is loosen, there will be more chance to detect the difference.

6.3 The Quantitative Relation between Power and the Four Elements

6.3.1 The test for single sample mean

Now let us derive the relation between the power and the four elements quantitatively for the test of (6.1) and the statistics (6.2) through any plots of Figs. 6.1–6.3. Say, (a) of Fig. 6.3 is used by regarding α_1 and β_1 as α and β . First of all, the upper plot shows that the dividing point is located at $\mu_0 + Z_\alpha\sigma/\sqrt{n}$ to ensure the upper tail equal to α , where Z_α is the upper critical value of the standard normal distribution corresponding to α . The lower plot shows, the dividing point divides the area under the density function of normal distribution $N(\mu_0 + \delta, \sigma^2/n)$ into two parts.

To the lower plot, after the distance between the dividing point and $\mu_0 + \delta$ being standardized, the dividing point is just the critical value of the standard normal distribution corresponding to the area of the upper tail equal to $1 - \beta$

$$\frac{(\mu_0 + Z_\alpha\sigma/\sqrt{n}) - (\mu_0 + \delta)}{\sigma/\sqrt{n}} = Z_{1-\beta}. \quad (6.11)$$

Hence

$$Z_{1-\beta} = \frac{-\sqrt{n}\delta}{\sigma} + Z_\alpha. \quad (6.12)$$

Notice that when $Z_{1-\beta}$ is smaller, $1 - \beta$ will be larger and when Z_α is smaller, α will be larger. One can see, (6.12) is consistent with the conclusion obtained from the qualitative discussion, that when δ is larger, σ is smaller, n is larger and α is larger, $Z_{1-\beta}$ will be smaller, hence the power $1 - \beta$ will be larger.

Although the above formula (6.12) is derived for a one-side test, it in fact also holds for a two-side test, as long as Z_α is regarded as the two-side critical value of the standard normal distribution corresponding to α . In practice, if a t test is applied for small sample, (6.12) can still be used to have a slightly optimal estimate for the power.

The first application of the formula (6.12) is to evaluate the power of a test with known values of δ , σ , n and α . The values of δ and σ , which are usually not available in practice, have to be estimated based on professional knowledge or pilot study.

Example 6.1 The length of effective period of a medication was initially six hours on average. It is said recently that the new product has prolonged the effectiveness seven hours. However, on the basis of 25 observed cases, a hypothesis test results in a non-significant outcome ($P > 0.05$) such that one cannot conclude the new product having the effectiveness longer than six hours on average. What is the problem?

Solution The hypothesis test is

$$H_0 : \mu = 6, \quad H_1 : \mu > 6.$$

The statistics is

$$Z = \frac{\bar{X} - 6}{\sigma / \sqrt{n}},$$

where $n = 25$. To evaluate the power of this test, the values of δ and σ are needed. Obviously, $\delta = 1$. The value of σ can be estimated through the 25 observed cases and, say, according to the former experience, $\sigma = 2$. Given $\alpha = 0.05$, the one-side $Z_{0.05} = 1.64$. Put all these into (6.12), we have

$$Z_{1-\beta} = \frac{-\sqrt{n}\delta}{\sigma} + Z_\alpha = \frac{-\sqrt{25} \times 1}{2} + Z_{0.05} = -0.86.$$

Check the table for standard normal distribution with $Z = -0.86$

$$1 - \beta = 0.8051.$$

It indicates that the power of the test is only 80.51%. In other words, if the new product really has the effectiveness prolonged to seven hours, then it may be missed by chance of 19.49%.

In order to promote the power of the test, the only choice is to increase the sample size.

6.3.2 The test for two sample means

For convenience, let both of the sample sizes in two groups equal to n , and assume the two random variables follow normal distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ respectively, where the variances of the two populations are assumed equal and known. To test

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 > \mu_2 \quad (6.13)$$

or

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 > 0 \quad (6.14)$$

the statistics to be used is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma \sqrt{2/n}}. \quad (6.15)$$

Given α , the decision rule is:

When $Z \geq Z_\alpha$, reject H_0 ; otherwise, not reject H_0 .

The condition $Z \geq Z_\alpha$ can be rewritten as

$$\bar{X}_1 - \bar{X}_2 \geq Z_\alpha \sigma \sqrt{2/n}.$$

For convenience, the alternative hypothesis in (6.14) can be written as

$$H_1 : \mu_1 - \mu_2 = \delta. \quad (6.16)$$

The upper plot of Fig. 6.4 is the density function of $\bar{X}_1 - \bar{X}_2$, $N(0, 2\sigma^2/n)$, when H_0 is true; the lower plot of Fig. 6.4 is the density function of $\bar{X}_1 - \bar{X}_2$, $N(\delta, 2\sigma^2/n)$, when H_1 is true; the vertical line stands on the dividing point $Z_\alpha \sigma \sqrt{2/n}$, where Z_α is still the upper critical value of the standard normal distribution corresponding to α . The area of the part with shadow in the upper plot is α , the area of the part without shadow in the lower plot is the power $1 - \beta$. To the lower plot, after the distance between the dividing point and δ being standardized, the dividing point is just the

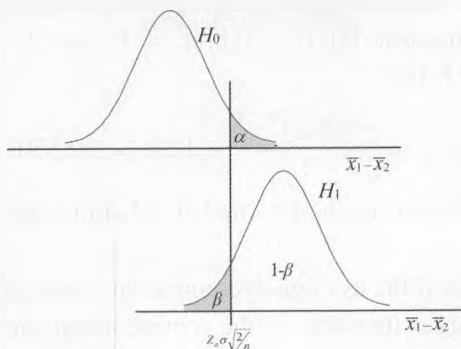


Fig. 6.4 A demonstration on the test for comparison between two means.

critical value of the standard normal distribution corresponding to the area of upper tail equal to $1 - \beta$,

$$\frac{Z_{\alpha}\sigma\sqrt{2/n} - \delta}{\sigma\sqrt{2/n}} = Z_{1-\beta} \quad (6.17)$$

or

$$Z_{1-\beta} = \frac{-\sqrt{n/2}\delta}{\sigma} + Z_{\alpha}. \quad (6.18)$$

Similar to (6.12), (6.18) shows, when δ is larger, σ is smaller, n is larger and α is larger, $Z_{1-\beta}$ will be smaller, hence the power $1 - \beta$ will be larger. In fact, (6.18) also holds for a two-side test, as long as Z_{α} is regarded as the two-side critical value of the standard normal distribution corresponding to α . In practice, if a t test is applied for small sample, (6.18) can still be used to have a slightly optimal estimate for the power.

Example 6.2 In a clinical trial on a hypotensor, there were two random samples with 15 cases respectively who were similar in disease condition and other important factors. One group was treated with the routine medication and another was treated with a newly developed medication. After a t test with $\alpha = 0.05$, the difference of the effectiveness between two groups were not significant statistically. How to understand such a matter? (Comparing to the routine medication, the new medication is not worthwhile to be applied in clinic unless the difference of the average decrement in blood pressure is larger than 0.8 kPa; by experience, the variance of the decrement in blood pressure within each group is about 1 kPa.)

Solution Since the difference was not significant statistically, one had to see the power of the test. Put $\delta = 0.8$, $\sigma = 1$, $n = 15$ and the one-side $Z_{0.05} = 1.64$ into (6.18),

$$Z_{1-\beta} = \frac{-0.8\sqrt{15/2}}{1} + 1.64 = -0.5509.$$

By checking the table of standard normal distribution with -0.5509 , $1 - \beta = 0.7088$.

One can see that if the average decrement on blood pressure of the new medication was larger than that of the routine medication by 0.8 kPa, the test just had a chance of 70.88% to detect it. The key was that the sample size was too small.

6.3.3 The test for two frequencies (large sample)

Assume that there are two groups of sample, $n_1 = n_2 = n$, the occurrence frequencies of a specified event are P_1 and P_2 , and the probabilities of that in the population are π_1 and π_2 . To test

$$H_0 : \pi_1 = \pi_2, \quad H_1 : \pi_1 > \pi_2. \quad (6.19)$$

The above-mentioned test for two group means can be applied on the basis of normal approximation for large samples, when the statistic is similar to (6.15), but σ should be replaced by

$$\sqrt{\frac{1}{2}[\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)]}.$$

Hence (6.18) becomes

$$Z_{1-\beta} = \frac{-\delta\sqrt{n}}{\sqrt{\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)}} + Z_\alpha, \quad (6.20)$$

where $\delta = \pi_1 - \pi_2$.

Similar to (6.18), (6.20) shows, when δ is larger, π_1 and π_2 are close to 0.5, n is larger and α is larger, $Z_{1-\beta}$ will be smaller, hence the power $1 - \beta$ will be larger. In fact, (6.20) also holds for a two-side test $H_1 : \pi_1 \neq \pi_2$, as long as Z_α is regarded as the two-side critical value of the standard normal distribution corresponding to α .

Example 6.3 Two groups of normal adults, 30 each, were randomly selected for a research project on the preventive effect of vitamin C. Group A was treated with vitamin C, and Group B was treated with placebo. The frequencies of bad cold were observed and compared for a fixed period of time. As a result, there were three and six persons getting bad cold during this period respectively. A test with $\alpha = 0.05$ resulted in that the difference in frequencies was not significant statistically. How to understand such an outcome?

Solution By experience of the researcher, the probability of getting a bad cold was about $\pi_1 = 20\%$. Assume vitamin C may reduce the probability to $\pi_2 = 10\%$ or even less. Put the estimated values of π_1 , π_2 , $n = 30$ and

$\alpha = 0.05$ into (6.20),

$$Z_{1-\beta} = \frac{-0.10\sqrt{30}}{\sqrt{0.20(1-0.20) + 0.10(1-0.10)}} + 1.645 = 0.5446.$$

Hence $1 - \beta = 0.2929$.

One can see that the power of this study is only 29.29%. In other words, if vitamin C were able to reduce the probability of getting bad cold to $\pi_2 = 10\%$, the chance to find such a difference by this study was only 29.29%, and there was about 70% chance to report a statistically non-significant result.

6.4 Estimation of Sample Size for the Tests in Common Use

The estimation of sample size is just the opposite of the calculation for power, that is, they share the same formula, but exchange the position of the known and unknown variables.

6.4.1 The test for single sample mean

To test for one group mean, we have had a formula in (6.12), which has been applied in Example 6.1 to evaluate the power of a test. Now rewrite (6.12), we have

$$n = \left(\frac{Z_\alpha + Z_\beta}{\delta/\sigma} \right)^2, \quad (6.21)$$

where $Z_\beta = -Z_{1-\beta}$. Then, given the specified values of α and β , and an appropriate guess of δ/σ , the sample size n can be easily estimated.

Example 6.4 To well demonstrate the new medication in Example 6.1, what should the least sample size be?

Solution In order to make the chance of claiming a trivial medication as an excellent one less than 5%, take $\alpha = 0.05$; in order to make the chance of claiming an excellent medication as a trivial one less than 1%, take $\beta = 0.01$. According to Example 6.1, keep $\delta = 1$ and $\sigma = 2$. From the table for standard normal distribution, the one-side $Z_{0.05} = 1.64$ and

$Z_{0.01} = 2.33$. Put all these values into (6.21), we have

$$n = \left(\frac{Z_{0.05} + Z_{0.01}}{1/2} \right)^2 = \left(\frac{1.64 + 2.33}{0.5} \right)^2 = 63.0436 \approx 63.$$

This shows, if the difference between the two means is one hour, to expect a one-side test at a level of $\alpha = 0.05$ to identify such a difference with a probability 99%, the sample size should at least be 63. Obviously, if the difference between the two means is greater than one hour, the power under such a sample size will be higher than 99%; contrarily, if the difference between two means is less than one hour, the power under such a sample size will not be able to reach 99%.

In addition, (6.21) is based on the condition of known σ . When σ is unknown, it could be replaced by the sample standard deviation S . However, the revised statistic in (6.2) will follow a t distribution when H_0 is true; Z_α should be replaced by t_α , the critical value of t distribution; and Z_β should be replaced by t_β^* , which is the critical value of a non-central t distribution. All of them depend on the sample size n . Then we are falling in a loop: the sample size is unknown, how can we get t_α and t_β^* ? In theory, one can put Z_α and Z_β into (6.21) to get a value for n ; then get t_α and t_β^* by degrees of freedom $n - 1$; put t_α and t_β^* into (6.21) again to get an updated n ; ... in such a way, by iteration, one may get a final result for the sample size. In practice, it is not necessary to be so fancy. By experience, one may get n by Z_α and Z_β first, and then add an extra term $0.5Z_\alpha^2$ to it. Say, the sample size for Example 6.3 could be

$$n = 63.0436 + 0.5(1.64)^2 = 64.3884 \approx 65.$$

6.4.2 The test for two sample means

Similarly, rewrite (6.18), we have

$$n = 2 \left(\frac{Z_\alpha + Z_\beta}{\delta/\sigma} \right)^2. \quad (6.22)$$

This looks similar to (6.21), but the sample size is doubled. It is because the variance of the difference between the two sample means is doubled.

In case that the value of σ comes from a pilot study with small sample, then the sample size needs to be added with an extra term $0.25Z_\alpha^2$.

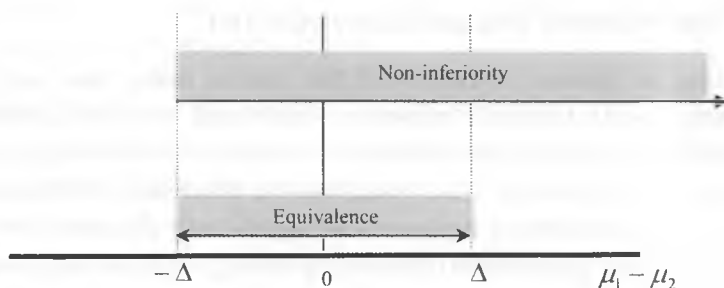


Fig. 6.5 Non-inferiority test and equivalence test.

If $p < \alpha$, reject H_0 , then we could conclude that new medication group is non-inferior to the standard group. When the new medication group is worse than the standard group, and the difference is over Δ , the probability to make the false conclusion of “new medication is non-inferior to the standard one” is less than α (Fig. 6.5).

Suppose the data come from a large sample of a normal distribution; $d = \bar{X}_1 - \bar{X}_2$ is the difference between the mean of new medication group and the mean of the standard group; S_d is the standard error of d , then the statistic for the non-inferiority test

$$Z = \frac{d + \Delta}{S_d}$$

follows the standard normal distribution when the null hypothesis holds. If the p -value is less than α , then we could conclude that the new medication is non-inferior to the standard one with a significance level α .

We can also use the confidence interval for a non-inferiority test:

- (1) To calculate the $(1 - \alpha)$ confidence interval of $\mu_1 - \mu_2$ as (a, b) ;
- (2) If the lower boundary a is larger than $-\Delta$, then we could also conclude that the new medication is non-inferior to the standard one with a significance level α .

6.5.2 Equivalence test

The equivalence test is to test whether the effect of the new medication is equivalent to that of the standard medication. Similar to the non-inferiority test, we need to determine a positive number Δ as the margin according

to the professional knowledge at first. After that, we need to conduct two non-inferiority tests (with significance level $\alpha/2$ for each):

$$H_{10} : \mu_1 - \mu_2 \leq -\Delta, \quad H_{1a} : \mu_1 - \mu_2 > -\Delta \quad (6.25a)$$

(whether the new medication is non-inferior to the standard one)

and

$$H_{20} : \mu_2 - \mu_1 \leq -\Delta, \quad H_{2a} : \mu_2 - \mu_1 > -\Delta \quad (6.25b)$$

(whether the standard medication is non-inferior to the new one).

If both H_{10} and H_{20} are rejected, both H_{1a} and H_{2a} are accepted. Then we could conclude that the new medication and the standard medication are equivalent. When the difference between the two medications is over Δ (including smaller than $-\Delta$ and larger than Δ), the probability to make the false conclusion of “the new medication and the standard medication are equivalent” is less than α (Fig. 6.5).

Suppose the data come from a large sample of a normal distribution; $d = \bar{X}_1 - \bar{X}_2$ is the difference between the mean of new medication group and the mean of the standard group; S_d is the standard error of d ; then both statistics for the equivalence test

$$Z_1 = \frac{d + \Delta}{S_d} \quad \text{and} \quad Z_2 = \frac{d - \Delta}{S_d} \quad (6.26)$$

follow the standard normal distribution when both of the null hypotheses hold. If both of the p -values are less than α , then we could conclude that the new medication is equivalent to the standard one with a significance level α .

We can also use the confidence interval for an equivalence test:

- (1) To calculate the $(1 - \alpha)$ confidence interval of $\mu_1 - \mu_2$ as (a, b) ;
- (2) If the lower and upper boundaries is in between $(-\Delta, \Delta)$, then we could also conclude that the two medications are equivalent with a significance level α .

Example 6.7 There was a trial to test whether the effects of angiotensin II receptor antagonists (experimental group, T) was non-inferior to the standard medication of angiotensin converting enzyme inhibitors (control group,

C) in treating mild to moderate hypertension. 240 subjects who met the inclusion criteria were randomly sampled and allocated to T or C group. The main outcome was supine diastolic blood pressure (SDBP, mmHg). The results showed that SDBP decreased 14 mmHg and 12 mmHg in the T and C groups, respectively. The difference value $d = 2$ mmHg, and the $S_d = 1.033$ mmHg. Previous clinical trials reported the minimal average reduction of SDBP was 10 mmHg. According to the clinical and statistical concern, we used $\Delta = 3$ mmHg (30% of effect size) as the margin for non-inferiority test.

Solution

- (1) One-side hypothesis testing:

Establishing the hypothesis:

$$H_0 : \mu_1 - \mu_2 \leq -3, \quad H_1 : \mu_1 - \mu_2 > -3; \quad \alpha = 0.05.$$

Calculating the statistic:

$$Z = \frac{d + \Delta}{S_d} = \frac{2 + 3}{1.033} = 4.84.$$

Since

$$Z = 4.84 > Z_{0.05} = 1.645, \quad p < 0.05.$$

Making conclusion:

H_0 was rejected, the effect of angiotensin II receptor antagonists was non-inferior to the standard medication of angiotensin converting enzyme inhibitors in treating mild to moderate hypertension.

- (2) Confidence interval:

The lower boundary of the confidence interval was:

$$2 - 1.645 \times 1.033 = 0.301 > -3.$$

Therefore, the conclusion was the same as the hypothesis testing that the effect of angiotensin II receptor antagonists was non-inferior to that of angiotensin converting enzyme inhibitors in treating mild to moderate hypertension.

If we need to conduct an equivalence test for this example, two one-side hypothesis tests as below should be performed:

Establishing the hypotheses:

$$H_{10} : \mu_1 - \mu_2 \leq -3, \quad H_{1a} : \mu_1 - \mu_2 > -3, \quad \alpha = 0.025.$$

$$H_{20} : \mu_2 - \mu_1 \leq -3, \quad H_{2a} : \mu_2 - \mu_1 < -3, \quad \alpha = 0.025.$$

Calculating the statistic:

$$Z_1 = \frac{d + \Delta}{S_d} = \frac{2 + 3}{1.033} = 4.84,$$

$$Z_2 = \frac{\Delta - d}{S_d} = \frac{3 - 2}{1.033} = 0.97.$$

Since $Z_1 = 4.84 > Z_{0.025} = 1.96$, therefore $p_1 < 0.025$; whereas, $Z_2 = 0.97 < Z_{0.025} = 1.96$, $p_2 > 0.025$.

Making conclusion:

H_{10} is rejected, the effects of angiotensin II receptor antagonists was non-inferior to that of angiotensin converting enzyme inhibitors on treating mild to moderate hypertension; However, H_{20} could not be rejected and we could not conclude that the effect of standard medication of angiotensin converting enzyme inhibitors was non-inferior to that of angiotensin II receptor antagonists on treating mild to moderate hypertension. Therefore, the equivalence test has failed.

6.6 Permutation Test¹

In Chap. 4, we know that t test is used for small samples of data which follow the normal distributions. However, if we use t test for the data which do not follow a normal distribution, the power of the t test will decrease.

The permutation test is more flexible than t test when we have insufficient information about the distribution of the data. Statistical inference is carried out based on the data characteristics and the logic of traditional hypothesis testing. Fisher had proposed the idea of permutation test in the 1930s, but no enough attention was paid due to the complicated calculation. With the development of the statistical softwares, SAS, S-Plus and R etc. can be used

¹We refer to Hesterberg T, Moore DS, Monaghan S, Clipson A, Epstein R. *Bootstrap Methods and Permutation Tests*, 2nd edn. New York: W. H. Freeman, 2005, for this section.

to implement this new approach. Nowadays the permutation test is widely applied in the multi-center clinical trials and genetic researches.

Now we take the comparison between two population means by independent samples as the example to interpret the principle of permutation test.

- (1) Establish the hypotheses:

H_0 : Two population distributions are the same;

H_1 : Two population distributions are different;

$\alpha = 0.05$ (two-side test)

- (2) Choose a statistic that measures the effect, D .

In this example, we use the difference between two sample means as the statistic: $D = \bar{X}_1 - \bar{X}_2$;

- (3) Calculate the observed value of the statistic.

In this example, $D(Obs) = \bar{x}_1 - \bar{x}_2$.

- (4) Under the condition that H_0 is true (two samples are from the same population), we pooled the two samples together, randomly assign the individuals into two groups (with the sample sizes consistent with those of the original samples), and calculate the new sample means \bar{X}'_1, \bar{X}'_2 and the statistic $D' = \bar{X}'_1 - \bar{X}'_2$.

- (5) Repeat step (4) for n times, we can get n values of D' , of which the distribution is called a *permutation distribution*.

- (6) Calculate the p -value: under the condition that H_0 is true, the p -value is defined as the probability when the statistic D is equal to or greater than the actually observed $D(Obs)$:

$$p = P(|D| \geq |D(Obs)|) \approx \frac{\text{number}(|D'| \geq |D(Obs)|)}{n}, \quad (6.27)$$

where the denominator n is the repeated times of the permutation resampling, the numerator is the times of $|D'| \geq |D(Obs)|$ among the n permutation re-samples;

- (7) Making the conclusion according to the p -value.

Example 6.8 We would like to know whether the new “directed reading activities” improve the reading ability of elementary school students, as measured by the scores of their Degree of Reading Power (DRP)? A study assigned the fourth-grade students at random to either the new method (Treatment group, 21 students) or the traditional teaching methods (Control

Table 6.1 Degree of reading power scores for fourth-graders.

Group		Scores											
Treatment	24	61	59	46	43	53	43	44	52	43	57	49	
	58	67	62	57	56	33	71	49	54				
Control	42	33	46	37	62	20	43	41	10	42	53	48	
	55	19	17	55	37	85	26	54	60	28	42		

group, 23 students). The DRP scores at the end of the study are listed in Table 6.1:

The procedure of permutation test is as follows:

- (1) Pooled the two samples together, and use the simple random sampling method to select 21 of 44 from the mixed sample and assign them to the treatment group, the rest are assigned to the control group. This is the process of permutation resample.
- (2) Calculate the mean DRP re-sample score in each group, using the individual DRP scores in Table 6.1. The difference between these means is our statistic: $D = \bar{X}_{\text{treatment}} - \bar{X}_{\text{control}}$.
- (3) Repeat this re-sampling from the 44 students for 1000 times. The distribution of the statistic from these 1000 re-samples is the permutation distribution (Fig. 6.6).
- (4) Calculate the $D(\text{Obs}) = 51.476 - 41.522 = 9.954$.
- (5) Locate this value on the permutation distribution to get the p -value.

The differences between permutation test and t test are as follows:

- (1) Hypotheses: The hypotheses for the t test are $H_0 : \mu_1 - \mu_2 = 0$, $H_1 : \mu_1 - \mu_2 \neq 0$; Whereas the hypothesis of permutation test is whether two populations' distributions are the same.
- (2) Statistics: The t statistic in the t test comes from the difference of the sample means. We used the same statistic in the permutation test here, but we could also use other statistic, such as the 25% trimmed means. Therefore statistic for permutation test is more flexible.
- (3) The calculation of statistic: The t test statistic is based on standardizing the difference of means in a formula to get a statistic. The permutation

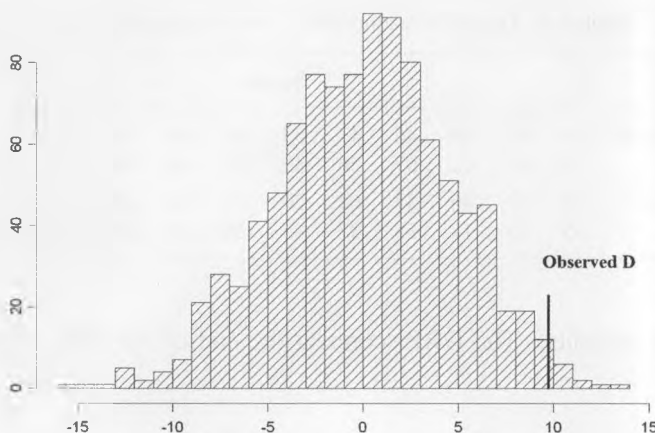


Fig. 6.6 The permutation distribution of the statistic $D = \bar{X}_{\text{treatment}} - \bar{X}_{\text{control}}$ based on the DRP scores of 44 students.

test directly gets the difference of means or some other statistics, and no formulas are necessary.

- (4) The calculation of p -values: The t test gives accurate p -values under the pre-condition of normal distribution. The permutation test gives approximate p -values even when the sampling distribution is not close to normal.

Permutation test is a “gold standard” for assessing two-sample t tests. If the two p -values of permutation test and t test differ considerably, it usually indicates that the pre-conditions for the method of two-sample t test are not suitable for these data.

6.7 Computerized Experiments

Experiment 6.1 On the power of t test for single sample mean Assume the average effect time of the medication in Example 6.1 following a normal distribution $N(7, 2^2)$. Randomly draw a sample from this population (sample size equals 25), and perform a t test at level $\alpha = 0.05$ to see whether one can conclude the population mean significantly different from 6 or not; repeat such a process for 100 times to count the number of times for correct conclusion and compare the percentage with the power 80.51% calculated in Example 6.1.

Program 6.1 Power of t test for single sample mean.

Line	Program	Line	Program
01	DATA TEST;	09	BY 1;
02	DO I=1 TO 100;	10	OUTPUT OUT=A
03	DO J=1 TO 25;		MEAN=MEAN STD=SD
04	X=7+2*RANNOR(0)-6;		STDERR=SE T=T PRT=P;
05	OUTPUT;	11	OPTIONS PAGESIZE=66;
06	END;END;	12	PROC PRINT DATA=A;
07	PROC MEANS NOPRINT;	13	WHERE P>0.10;
08	VAR X;	14	RUN;

In Program 6.1, line 04 is to generate a random number following $N(7, 2^2)$, and subtract 6 from it; lines 02–06 generate 100 random samples with sample size 25 for each; line 07, t test on whether the population mean equals 6 or not for the 100 samples respectively; lines 12 and 13 select the results of no rejection on the null hypothesis and print the corresponding sample.

Experiment 6.2 By changing the distribution in Experiment 6.1 to $N(8, 2^2)$, or changing the sample size to 49, or changing the α level to 0.2, repeat the experiment and compare the outcomes with that obtained from Experiment 6.1.

Experiment 6.3 Equivalence trials Example 6.7 is used as example. In Program 6.2, line 01, to construct a SAS data set named “equaltest”; lines 02

Program 6.2 Equivalence trials.

Line	Program	Line	Program
01	DATA EQUALTEST;	11	T2=(U-D)/SD;
02	NT=120;	12	P1=PROBT(-T1,
03	NC=120;		NT+NC-2);
04	MEANT=14;	13	P2=PROBT(-T2,
05	MEANC=12;		NT+NC-2);
06	L=-3;	14	PROC PRINT;
07	U=3;	15	VAR T1 P1;
08	D=MEANT-MEANC;	16	RUN;
09	SD=1.033;	17	PROC PRINT;
10	T1=(D-L)/SD;	18	VAR T2 P2;
		19	RUN;

Program 6.3 Permutation test.

Line	Program	Line	Program
01	DATA PERM;	07	PROC MULTTEST
02	INPUT X TRT@@;		PERMUTATION NSAMPLE=1000
03	CARDS;		SEED=23;
04	24 2 61 2 59 2 46 2 43 2 53 2	08	TEST MEAN(X);
05	...	09	CLASS TRT;
06	;	10	RUN;

and 03, to define the sample size of both experimental and control groups; lines 04 and 05 define the means of both experimental and control groups; lines 06 and 07 define the upper and lower limits; lines 10 and 11 calculate t_1 and t_2 ; lines 12 and 13 calculate the total amount of probability measure to the left of $-t_1$ and $-t_2$; lines 14–19 print the value of t_1 , p_1 , t_2 and p_2 .

Experiment 6.4 Permutation test Example 6.8 is used as an example. Lines 01–05 read the scores and groups of 44 students; lines 07–09, to conduct permutation test for two groups based on the seed 23, and re-sample 999 times.

6.8 Practice and Experiments

1. It is said that the average body temperature of normal people in a nation is higher than 37°C . To check, a sampling survey is planned. Suppose an increase of 0.1°C is significant; the standard deviation of body temperature of normal people is about 0.2°C ; the probabilities of type I and type II errors are limited by $\alpha = 0.05$ and $\beta = 0.05$. Estimate the necessary sample size, and confirm by computerized experiment.
2. Someone did not estimate the sample size for the above-mentioned survey and decided to take $n = 25$. Calculate the power and confirm it by computerized experiment.
3. To screen hypotensors, it is decided that only when its decrement of blood pressure is more than 2 kPa, the medication can be kept as one of the candidates for the next run of study. Now a pilot study for a medication has been completed for ten animals, as a result, the standard deviation of

the decrement of blood pressure is 0.5 kPa. What is the adequate sample size for the formal study?

4. To compare the dissolution rates of two types of tablets, randomly select ten tablets for each type and measure their dissolved amount in 5 minutes respectively, then perform a t test at a level of $\alpha = 0.05$. According to a pilot study, the standard deviation of both types of tablets is about 6 units, and the difference of their average 5-minute dissolution is also about 6 units. Evaluate the power of the above test. In order to reach a power of 95%, what is the adequate sample size?
5. The departments of inner medicine of two hospitals A and B have randomly selected 30 inpatients respectively, among which 20 in hospital A and 23 in hospital B are satisfied with the service they received. After a test, there is no statistically significant difference for satisfaction rate between two hospitals. In order to explore whether the difference on satisfaction rates between two hospitals is more than 10%, at least how many inpatients should be observed?
6. The above satisfaction rates in two samples are $20/30=67\%$ and $23/30=77\%$ respectively. The difference of them seems not small, but the statistical test can hardly reject the null hypothesis of equal satisfaction rates. Suppose the sample sizes were extended to 300 for each, and the sample satisfaction rates were still 67% and 77% respectively, then the same test resulted in a rejection of the null hypothesis. How to understand such a phenomenon? Explain from the point of power.
7. There is an equivalence trial to treat frequent urination, urgent urination, and urinary incontinence. The main outcome is the frequency of urination during the last 24 hours. The average reduced times of urination are a and b in the experimental group and control group, respectively. $\pm\Delta$ is defined as the threshold value. Conduct a hypothesis testing on this equivalence trial; and discuss under what conditions is the experimental group equivalent with the control group?
8. Let us perform a permutation test by hand for a small random subset of the DRP data (Example 6.8). Here are the data:
Treatment group: 24 61
Control group: 42 33 46 37

(1) Calculate the D : the actually observed difference value between two groups;

- (2) Resample: mix the two groups together, then simple randomly (use die etc.) select two persons as the treatment group, and the rest four persons are assigned to the control group. After that, calculate the difference value between two groups.
- (3) Repeat step (2) for 20 times to get the permutation distribution;
- (4) What proportions of the 20 statistics values are equal to or greater than the actually observed D ?

(1st edn. Jiqian Fang; 2nd edn. Chun Hao, Jiqian Fang, Xiaoyu Zuo)

Chapter 7

Single-Factor Analysis of Variance

Single factor design has only one treatment factor with G levels ($G \geq 2$). If there are no control factors, it is completely random design; if one control factor exists, it is randomized complete-block design; if there are two control factors, it is Latin-square design. The completely random design, randomized complete-block design and Latin-square design are the fundamental methods of experiment design, whose results data are usually analyzed by analysis of variance (ANOVA). This chapter is to introduce single-factor analysis of variance, and the multi-factor analysis of variance can be seen in Chap. 22.

7.1 One-Way Analysis of Variance: Completely Random Design

7.1.1 *The completely random design*

For the completely random design, there is only one treatment factor with $G(\geq 2)$ levels. The term *level* refers to the possible status planned for the treatment *factor*. In other words, a total of G treatment groups are planned in the experiment. For example, in a clinical trial the factor may be the drug, the treatments could be different drugs or different dosages of the same drug. N experimental units are randomly assigned into G treatment groups. N is known as sample size, which is the total number of units in G treatment groups. The sample sizes in different treatment groups are noted as n_1, n_2, \dots, n_G , which may or may not be equal. It is called a balanced design if sample sizes are equal and $N = Gn$; otherwise, it is called an unbalanced

Table 7.1 Randomized grouping result.

Unit number	1	2	3	4	5	6	7	8	9	10	11	12
Random number	39	90	22	00	66	82	89	08	92	72	36	60
Rank (R)	5	11	3	1	7	9	10	2	12	8	4	6
Grouping result	2	3	1	1	2	3	3	1	3	2	1	2

design. The efficiency of balanced design is better than that of unbalanced design given the same N . The completely random allocation usually results in an unbalanced design, which is the simplest and fundamental design. The main task of data analysis for this kind of experiments is to compare means of the G independent treatments.

Example 7.1 Randomly assign 12 laboratory blood specimens (experiment units) into three groups with four blood specimens in each group.

Solution Assign numbers 1 to 12 to the experiment units in a convenient manner, for instance, consecutively by certain characteristics of the experiment units such as the order of animal's weights or the order of patients visiting a clinic. In the example showed in Table 7.1, blood samples are given a number in the order they are collected; then select 12 two-digit random numbers from a random number table, a statistical calculator or a computer; sort the random numbers and rank them from 1 to 12 (the R 's in Table 7.1); assign blood specimens with $R = 1-4$ to group 1, $R = 5-8$ to Group 2, and $R = 9-12$ to group 3. Table 7.1 shows that units 3, 4, 8, and 7 have been assigned to group 1, etc.

Note:

1. The digits of the random number must be more than that of N , and skip those random numbers that appear more than once.
2. If the sample sizes required for different groups are proportional, then the allocation needs to be adjusted accordingly. In Example 7.1, suppose you need five blood specimens in group 1, four blood specimens in group two, and three blood specimens in group 3, you can assign $R = 1-5$ to group 1, $R = 6-9$ to group 2, and $R = 10-12$ to group 3. If N is bigger (say $N > 30$), computer software may be used to create random numbers and

get the ranks by sorting the random numbers. For instance, randomly allocate 50 patients into three groups, $n_1 = n_2 = 20$, $n_3 = 10$, one can create the ranks from 1 to 50, and assign the patients with rank 1 to 20 into the first group, those with rank 21 to 40 into the second group, and those with 41 to 50 into the third group.

7.1.2 The underlying concepts of ANOVA

Example 7.2 12 blood specimens are randomly assigned into three groups according to Table 7.1. Group 1 receives the treatment of anticoagulant A; group 2 receives anticoagulant B; and group 3 receives anticoagulant C. For each blood specimen, the erythrocyte sedimentation rate (ESR) after receiving the treatment is measured. The aim is to test whether the three mean ESRs are significantly different. The results are showed in Table 7.2.

There are three kinds of variations in Table 7.2. The first is the total sum of squared deviations from the overall mean, denoted as SS_T , which represents the deviation of any observation from the "grand mean" $\bar{X} = 12.2$. The second is the within-group sum of squared deviations, or error sum of squared deviations, calculated from the sum of squares of each observation about its own group mean, denoted as SS_E , which represents the deviation of any observation in each group from the "group mean" \bar{X}_i , 16.0, 11.3 and 9.3. The third is the between-group sum of squared deviations, calculated from the sum of squares of the deviations of each group mean about the grand mean, denoted as SS_B , which represents the deviation of group mean \bar{X}_i from the "grand mean" \bar{X} . The three kinds of the sum of squares

Table 7.2 Erythrocyte sedimentation rate (ESR mm/h).

Anti-coagulant	ESR (x_{ij})	Group means (\bar{x}_i)	Within group means (s_i^2)	Total mean (\bar{x})	Combined variance (s_c^2)
A	17 16 16 15	16.0	0.67	12.2	3.17
B	10 11 12 12	11.3	0.92		
C	11 9 8 9	9.3	1.58		

can be written as

$$\text{Total} \quad SS_T = \sum_{i=1}^G \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2. \quad (7.1)$$

$$\text{Between-group} \quad SS_B = \sum_{i=1}^G n_i (\bar{X}_i - \bar{X})^2. \quad (7.2)$$

$$\text{Within-group} \quad SS_E = \sum_{i=1}^G \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2. \quad (7.3)$$

For the above example,

$$\begin{aligned} SS_T &= \sum_{i=1}^G \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 \\ &= [(17 - 12.2)^2 + (16 - 12.2)^2 + (16 - 12.2)^2 + (15 - 12.2)^2] \\ &\quad + [(10 - 12.2)^2 + (11 - 12.2)^2 + (12 - 12.2)^2 + (12 - 12.2)^2] \\ &\quad + [(11 - 12.2)^2 + (8 - 12.2)^2 + (8 - 12.2)^2 + (9 - 12.2)^2] \\ &= 105.67, \end{aligned}$$

$$\begin{aligned} SS_B &= \sum_{i=1}^G n_i (\bar{X}_i - \bar{X})^2 \\ &= 4(16.0 - 12.2)^2 + 4(11.3 - 12.2)^2 + 4(9.3 - 12.2)^2 = 96.17, \end{aligned}$$

$$\begin{aligned} SS_E &= \sum_{i=1}^G \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 = SS_{EA} + SS_{EB} + SS_{EC} \\ &= [(17 - 16.0)^2 + 2 \times (16 - 16.0)^2 + (15 - 16.0)^2] \\ &\quad + [(10 - 11.3)^2 + (11 - 11.3)^2 + 2 \times (12 - 11.3)^2] \\ &\quad + [(11 - 9.3)^2 + (8 - 9.3)^2 + 2 \times (9 - 9.3)^2] = 9.50. \end{aligned}$$

One can see that $105.67 = 96.17 + 9.50$.

In general, it can be proved that

$$SS_T = SS_E + SS_B. \quad (7.4)$$

Denote ν as degree of freedom accordingly,

$$\nu_T = \nu_E + \nu_B. \quad (7.5)$$

Then the mean squares or variances will be

$$\text{Within-group} \quad MS_E = SS_E / \nu_E. \quad (7.6)$$

$$\text{Between-group} \quad MS_B = SS_B / \nu_B. \quad (7.7)$$

The basic calculation of analysis of variance is to divide the total variance into within-group variance and between-group variance, and compare the two variances by the test statistic F

$$F = \frac{MS_B}{MS_E}. \quad (7.8)$$

If F is greater than a given critical value at significant level α , it is reasonable to use it as evidence that differences exist and the differences are larger than what random variation can explain. If F is less than the critical value, it indicates different groups may be equally effective, and the differences among sample means may be resulted from random variation.

The above is the main idea of the analysis of variance. If $G = 2$, it is exactly the same as the two-sample t test. In fact, the value of t is the positive square root of the value of F , that is, $t = \sqrt{F}$ when $G = 2$.

7.1.3 Assumptions on analysis of variance and residual analysis

The analysis of variance of a completely random design is also called a one-way analysis of variance. The basic assumptions underlying one-way analysis of variance are:

- (1) Independence, that is, the measurement results X_{ij} are independent from one another;
- (2) Normal distribution, that is, X_{ij} follows normal distribution $N(\mu_i, \sigma_i^2)$, $i = 1, 2, \dots, G$;
- (3) Homogeneity of variances, $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_G^2$.

To investigate if results accord with the above assumptions, residual analysis can be performed after establishment of analysis of variance model.

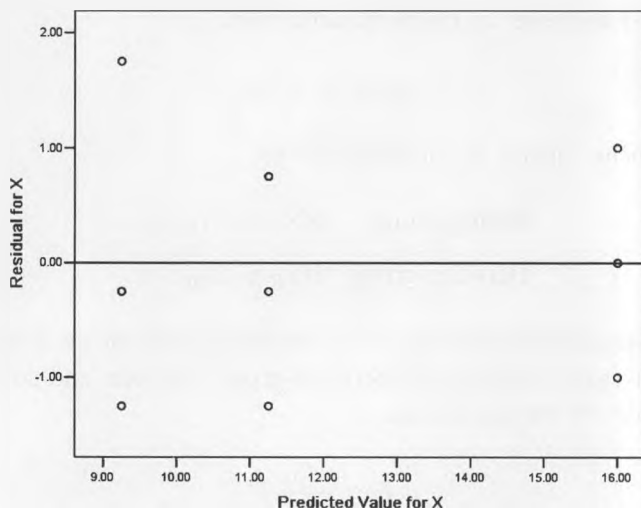


Fig. 7.1 Scatter plot for check of independence assumption.

For a completely random design, residual can be defined as:

$$e_{ij} = x_{ij} - \bar{x}_i \quad (7.9)$$

that is, difference between investigation values and group mean. The group mean is the predicted value. e_{ij} can be presented as plots of residuals in three modes for residual analysis.

(1) Residuals plot for independence assumption

Take the group mean as horizontal axis and residual as vertical axis to draw scatter plot. If the experiment results accord with the independence assumption, scatter plot is equally distributed above and below the horizontal line of zero, see Fig. 7.1.

(2) Residuals plot for check of normal distribution

Residuals can be drawn as $Q - Q$ plot ($Q - Q$ plot, see Chap. 2). If the experiment results are normally distributed, $Q - Q$ plot is almost a line, and residuals bigger or less than zero are equally distributed, see Fig. 7.2.

(3) Residuals plot for check of homogeneity of variances

To take the group number as horizontal axis (for example, three anti-coagulants are given the continual number of 1, 2, 3) and residual as vertical axis to draw scatter plot. If the experiment results accord with

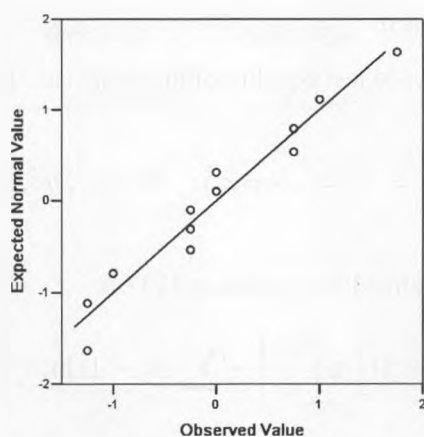


Fig. 7.2 $Q-Q$ plot for check of normal distribution.

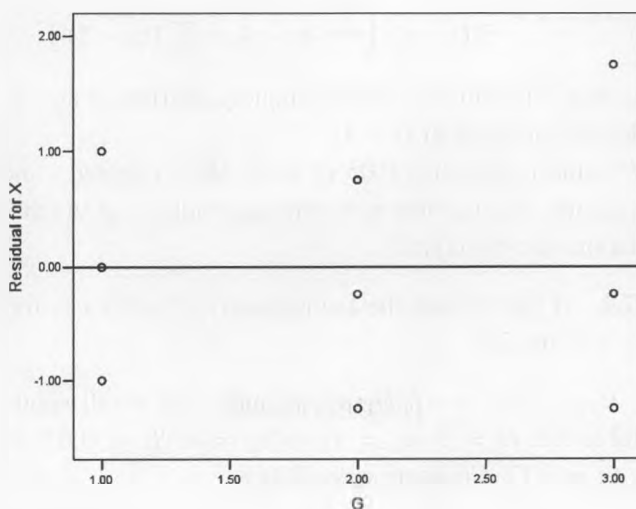


Fig. 7.3 Scatter plot for check of homogeneity of variances.

the homogeneity of variances assumption, the distribution of residuals for each group is similar, see Fig. 7.3.

In practice, the judgment of residuals is mainly dependent on the researchers' experience, especially when the number of sample in each group is less.

7.1.3.1 Bartlett's test

There are many ways to test equality of the variances. The commonly used is Bartlett's test

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_G^2, \quad H_1: \sigma_1^2, \sigma_2^2, \dots, \sigma_G^2$$

are not all equal.

A test statistic defined by Bartlett in 1937 is

$$\chi^2 = \frac{1}{m} \left\{ \left[\sum (n_i - 1) \right] \ln \left[\frac{S_c^2}{G} \right] - \sum (n_i - 1) \ln(S_i^2) \right\}, \quad \nu = G - 1, \quad (7.10)$$

where

$$m = 1 + \frac{1}{3(G-1)} \left[\sum \frac{1}{n_i - 1} - \frac{1}{\sum (n_i - 1)} \right].$$

When H_0 is true, this statistic is approximately distributed as a χ^2 distribution with degrees of freedom $G - 1$.

If the P -value is less than 0.05 or 0.10, H_0 is rejected; otherwise H_0 cannot be rejected, that is, there is not enough evidence to say the variances of G populations are not equal.

Example 7.2 (Cont'd) Test the homogeneity of variances for the three populations in Table 7.2.

Solution $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$, $H_1: \sigma_1^2, \sigma_2^2$ and σ_3^2 are not all equal $\alpha = 0.05$.

From Table 7.2, $G = 3$, $n_1 = n_2 = n_3 = 4$, $S_1^2 = 0.67$, $S_2^2 = 0.92$, $S_3^2 = 1.58$, $S_c^2 = 3.17$. Compute with (7.9)

$$m = 1 + \frac{1}{3(3-1)} \left[\sum \frac{1}{4-1} - \frac{1}{\sum (4-1)} \right] = 1.148,$$

$$\begin{aligned} \chi^2 &= \frac{1}{1.148} \left\{ \left[\sum (4-1) \right] \ln \left[\frac{3.17}{3} \right] \right. \\ &\quad \left. - (4-1)(\ln 0.67 + \ln 0.92 + \ln 1.58) \right\} \\ &= 0.50, \quad \nu = 3 - 1 = 2. \end{aligned}$$

Referring to χ^2 critical values in Table 7 of Appendix II, $\chi_{0.10,2}^2 = 4.61$, $P > 0.10$, H_0 cannot be rejected.

By experience, if the ratio of the maximal sample variance S_{\max}^2 to the minimal variance S_{\min}^2 among all the groups is greater than 2.5, one may consider that the variances are not equal without homogeneity test.

7.1.3.2 Test for transformations

If the experiment results are not accord with at least one of the assumptions through residual analysis, some transformations of the scale of measurement may improve the departures from normality as well as the heterogeneity of variances. The following transformations are often used:

$$Y = \log(X + a), \quad (7.11)$$

$$Y = \sqrt{X}, \quad (7.12)$$

$$Y = \sin^{-1} \sqrt{p}. \quad (7.13)$$

(7.11) is used for positive skew continuous variable, where a is a constant; (7.12) is used for Poisson variable; (7.13) is used for a variable of a proportion greater than 0 and less than 1 where \sin^{-1} is the function arcsine with a unit of radian.

7.1.4 Hypothesis test of means (one-way analysis of variance)

Example 7.2 (Cont'd) Table 7.2 shows three sample means about the effects of anticoagulant A, B and C, $\bar{X}_1 = 16.0$, $\bar{X}_2 = 11.3$, $\bar{X}_3 = 9.3$. Test if there are any differences among the population means of erythrocyte sedimentation rate (ESR) corresponding to the three anticoagulant A, B and C.

Solution It is a typical statistical inference to make a decision if the three anticoagulants A, B and C are differently effective to ESR for all blood specimens rather than just the 4 individuals in each group of the experiment. It is statistically reasonable to believe that the population means of ESR is the same before applying the three anticoagulants because the blood specimens were randomly assigned into three groups.

Table 7.3 The table for one-way analysis of variance.

Source	<i>DF</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Between groups	$G - 1$	SS_B	$MS_B = SS_B/(G - 1)$	MS_B/MS_E
Within groups (Errors)	$N - G$	$SS_E = SS_T - SS_B$	$MS_E = SS_E/(N - G)$	
Total	$N - 1$	SS_T		

Let μ_1, μ_2 and μ_3 be the population means after the treatments. Now we would test

$$H_0 : \mu_1 = \mu_2 = \mu_3, \quad H_1 : \mu_1, \mu_2 \text{ and } \mu_3 \text{ are not all the same,} \\ \alpha = 0.05.$$

The values of F and degrees of freedom can be calculated according to Table 7.3 for one-way analysis of variance. There the SS_B is the same as that in (7.3), but for convenience of calculation, (7.3) could be written as

In Table 7.3, source means the source of variance, DF means degrees of freedom, SS means the sum of squares of the deviations, MS means the mean squares of deviations,

Based on the data in Table 7.2, an ANOVA table can be obtained.

$$SS_T = \sum_{i=1}^3 \sum_{j=1}^4 (X_{ij} - 12.2)^2 = 105.67,$$

$$SS_B = \sum_{i=1}^G n_i (\bar{X}_i - \bar{X})^2 \\ = 4 \times (16.0 - 12.2)^2 + 4 \times (11.3 - 12.2)^2 + 4 \times (9.3 - 12.2)^2 \\ = 96.17,$$

$$SS_E = SS_T - SS_B = 9.50,$$

$$\nu_B = 3 - 1 = 2, \quad MS_B = 96.17/2 = 48.09,$$

$$\nu_E = 12 - 3 = 9, \quad MS_E = 9.50/9 = 1.06,$$

$$F = 48.09/1.0 = 45.37.$$

Table 7.4 Table of one-way ANOVA for the effects of the anticoagulants.

Source	<i>DF</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>P</i>
Between the anticoagulants	2	96.17	48.09	45.37	<0.05
Errors	9	9.50	1.06		
Total	11	105.67			

Table 7.5 Defervescence time of patients with type-B encephalitis in three groups (days).

Groups	Defervescence time (x_{ij})	\bar{x}	s_c^2	\bar{x}_i	s_i^2
A	0, 2, 0, 0, 5, 9		2.67	13.4667	
B	32, 13, 6, 7, 10, 2	7.05	55.30	11.6667	113.0667
C	0, 11, 15, 11, 3, 1		6.83	39.3667	

The results are listed in Table 7.4. Referring to F critical values in Table 6.1 of Appendix II, $F_{0.05,2,9} = 4.26$. Since $45.37 > F_{0.05,2,9}$, $P < 0.05$, H_0 is rejected and alternative hypothesis H_1 is accepted. That is to say that there may exist differences among different population means.

Example 7.3 18 patients with acute encephalitis B in a clinic were randomly allocated into three groups. Each group accepted different treatment, say treatment A, treatment B and treatment C; and the days with fever were measured as the effects of treatments, see Table 7.5. Make an inference from the differences of means of days fever with among the three groups whether the treatments had different effects.

Solution (1) Calculate descriptive statistical values, see Table 7.5.

(2) Residual plot for check of independence, see Fig. 7.4.

(3) Residuals plot for check of normal distribution, see Fig. 7.5.

(4) Residuals plot for check of homogeneity of variances, see Fig. 7.6.

(5) Test for transformations

The residuals in Fig. 7.4 showed sector distribution, that is, the bigger the group means, the bigger the residuals, which does not accord with the independence assumption. The $Q-Q$ plot of residuals in Fig. 7.5

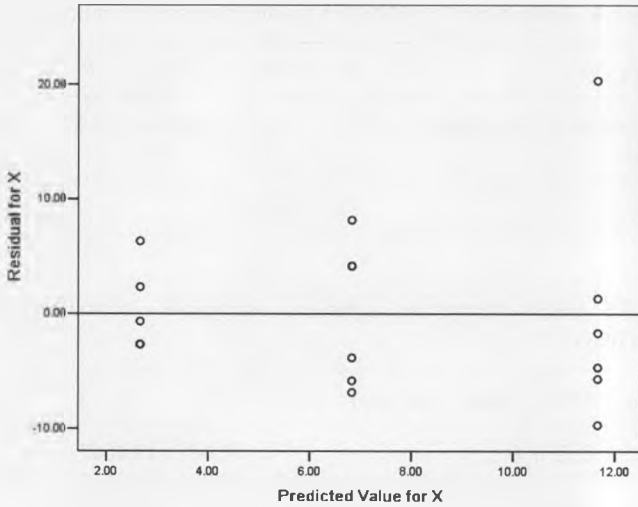


Fig. 7.4 Scatter plot for check of independence assumption for Table 7.5.

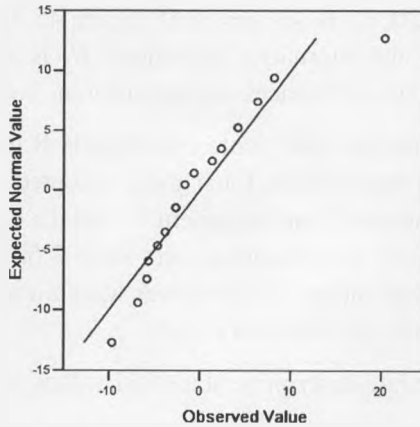


Fig. 7.5 Scatter plot for check of normal distribution for Table 7.5.

showed obvious curvilinear relationship, which does not accord with the assumption of normal distribution (for professional consideration, deferescence time may be positive skewness distribution). In Fig. 7.6, residuals in groups A and C showed a distribution between -10 to 10 , however, residuals in group B was between -10 to 20 , which might not accord with the assumption of normal distribution. Figure 7.6 also showed positive

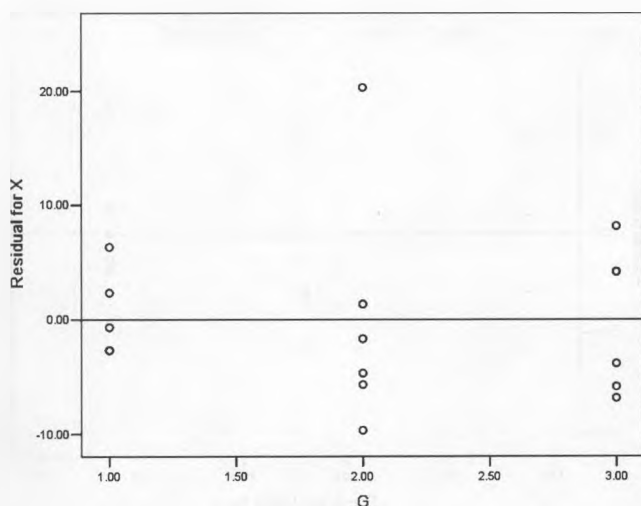


Fig. 7.6 Scatter plot for check of homogeneity of variances for Table 7.5.

Table 7.6 Square root transformation for data in Table 7.5.

Treatment groups	$y = \sqrt{x}$	\bar{y}	s_c^2	\bar{y}_i	s_i^2
A	0, 1.414, 0, 0, 2.236, 3		1.11	1.73	
B	5.657, 3.606, 2.449, 2.646, 3.162, 1.414	2.16	2.05	3.16	2.05
C	0, 3.317, 3.873, 3.317, 1.732, 1		2.21	2.36	

proportion between group mean and variance. The maximum variance 113.0667 was approximately 10 times of minimum variance 13.4667. Through Bartlett's test, $\chi^2 = 4.76$, $P < 0.10$, it showed heterogeneity of variance. In Table 7.6, square-root transformation in (7.10) was performed, and the residuals plots are in after transformation are in Figs. 7.7–7.9.

After transformation, sector distribution of residuals was obviously improved (Fig. 7.7); the scatter in $Q - Q$ plot was equally distributed (Fig. 7.8); the distribution ranges for group A, B and C were similar (Fig. 7.9). The ANOVA table can be obtained by putting description statistics from Table 7.6 into Table 7.3 (Table 7.7).

Conclusion: At the level of $\alpha = 0.05$, there has no statistical significance for the difference among the three groups.

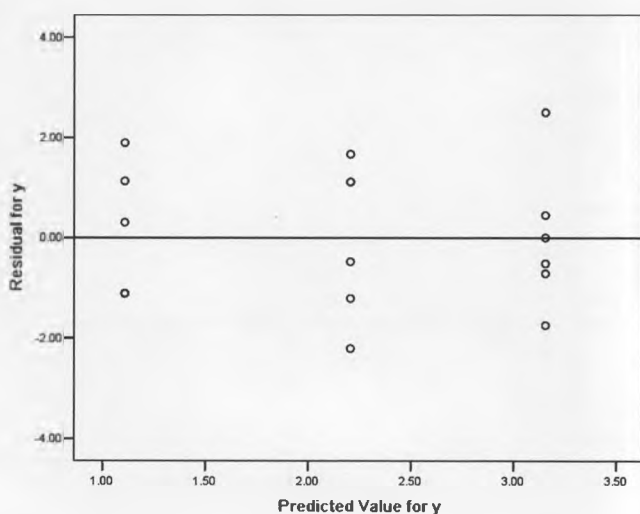


Fig. 7.7 Scatter plot for check of independence assumption for Table 7.6.

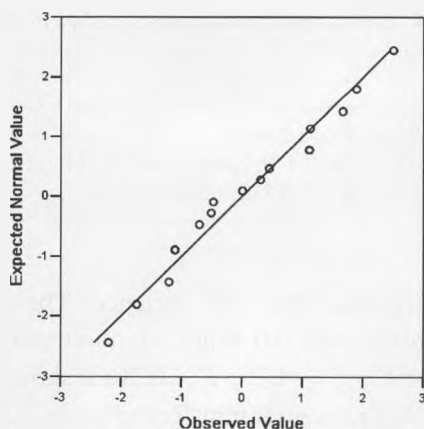


Fig. 7.8 Scatter plot for check of normal distribution for Table 7.6.

7.1.5 Multiple comparisons

When $G > 2$ and the null hypothesis $H_0 : \mu_1 = \mu_2 = \dots = \mu_G$ is rejected by a one-way ANOVA, the experimenters still do not know whether the differences of means between any pair of groups exist? Multiple comparisons will solve the problem and sometimes do not rely on the result of

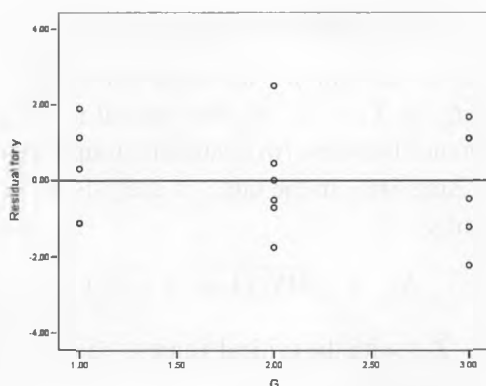


Fig. 7.9 Scatter plot for check of homogeneity of variances for Table 7.6.

Table 7.7 Analysis of variance for data in Table 7.6.

Source	<i>DF</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>P</i>
Between groups	2	12.5965	6.2983	3.08	>0.05
Errors	15	30.6700	2.0447		
Total	17	43.2665			

ANOVA table. There are at least two situations to form the null hypotheses for multiple comparisons.

1. To examine whether two specified means are equal or not. The null hypothesis is $H_0 : \mu_i = \mu_j (i \neq j)$. The probability of type I error when H_0 is rejected is called comparison-wise error rate (CER). The tests can be performed without doing ANOVA in advance.
2. To examine whether all the means of comparison groups are equal. $H_0 : \mu_i = \mu_j (i < j, i, j = 1, 2, \dots, G)$. Usually the tests are performed after rejecting H_0 by the analysis of variance to show more details of differences between every pairs of means. The probability of type I error in rejecting H_0 is called experiment-wise error rate (EER).

There are various methods of multiple comparisons in statistics. Here, some frequently used methods are introduced which are used to control CER and EER.

7.1.5.1 LSD-t test (least significant difference t test)

Make an inference if μ_i and μ_j are different based on the selected experiment result $\bar{d}_{ij} = \bar{X}_i - \bar{X}_j$. \bar{d}_{ij} has special meaning in the experiment, e.g. the difference between j th treatment group and the control group with treatment i . Using MS_E in the table of analysis of variance, compute the standard error of \bar{d}_{ij}

$$S_{\bar{d}_{ij}} = \sqrt{MS_E(1/n_i + 1/n_j)} \quad (7.14)$$

and compare $|\bar{X}_i - \bar{X}_j|$ with the critical value $t_{\alpha, \nu} S_{\bar{d}_{ij}}$.

$H_0: \mu_i = \mu_j$ is rejected if

$$|\bar{X}_i - \bar{X}_j| \geq t_{\alpha, \nu} S_{\bar{d}_{ij}}, \quad (7.15)$$

where $t_{\alpha, \nu}$ is the critical value of t distribution with degrees of freedom ν , which is the DF of error or DF within groups in the table of analysis of variance.

Example 7.2 (Cont'd) For the anticoagulant experiment, $\bar{X}_1 = 16.0$, $\bar{X}_2 = 11.3$, $\bar{X}_3 = 9.3$ are listed in Table 7.2, and $\nu=9$, $MS_E = 1.06$ have been presented in Table 7.4. Compare the effect of anticoagulant A with B and A with C .

Solution $n_1 = n_2 = n_3 = 4$, $MS_E = 1.06$. Compute with (7.13), $S_{\bar{d}_{12}} = S_{\bar{d}_{13}} = 0.73$. Given $\alpha = 0.01$, $\nu = 9$, referring Table 5 of Appendix II, $t_{0.03, 9} = 3.25$. Compute with (7.14), $S_{\bar{d}_{ij}} = 3.25 \times 0.73 = 2.37$.

$H_0: \mu_1 = \mu_2$, $|\bar{X}_1 - \bar{X}_2| = 16.0 - 11.36 = 4.70 > 2.37$, $P < 0.01$, reject H_0 .

$H_0: \mu_1 = \mu_3$, $|\bar{X}_2 - \bar{X}_3| = 11.3 - 9.3 = 2.00 < 2.37$, $P > 0.01$, do not reject H_0 .

It is concluded that the effects of anticoagulant A and anticoagulant B are different at significant level $\alpha = 0.01$. But that of anticoagulant B and anticoagulant C are not statistically different.

7.1.5.2 SNK-q test (Student-Newman-Keuls q test)

Make all possible comparisons among all group means in an experiment after the analysis of variance rejecting the null hypothesis $H_0: \mu_1 = \mu_2 = \dots = \mu_G$. Using MS_E in the table of analysis of variance,

compute the standard error of mean for each group if $n_1 = n_2 = \dots = n_G = n$,

$$S_{\bar{X}} = \sqrt{MS_E/n}. \quad (7.16)$$

If n 's are not equal, substitute n into (7.15) with

$$n = \frac{1}{G-1} \left(N - \frac{\sum n_i^2}{N} \right).$$

All means should be sorted from the smallest to the biggest to form contrasts. Each contrast may contain a means, $a = 2, 3, \dots, G$. In Example 7.2, the means of three groups are sorted as 9.3, 11.3 and 16.0; if 9.3 and 11.3 are selected to form a contrast, $a = 2$; if 9.3 and 16.0 are contrasted, $a = 3$, because 11.3 is included in the range and the contrast 9.3–16.0 contains three means; if 11.3 and 16.0 are contrasted, $a = 2$ as well. With the parameters a and v , the critical value $q_{a,a,v}$ of SNK- q test can be found from Table 7 of Appendix II. $H_0 : \mu_i = \mu_j$ is rejected if

$$|\bar{X}_i - \bar{X}_j| \geq q_{a,a,v} S_{\bar{X}_{ij}}. \quad (7.17)$$

Example 7.2 (Cont'd) From Table 7.4, $H_0 : \mu_1 = \mu_2 = \mu_3$ was rejected by the analysis of variance. That is to say that there were at least two group means different among the three. Do all comparisons among means to find out if any difference exists among the anticoagulants.

Solution From Table 7.2, $\bar{X}_1 = 16.0$, $\bar{X}_2 = 11.3$, $\bar{X}_3 = 9.3$. And from Table 7.4, $v = 9$, $MS_E = 1.06$. Three comparisons may be performed, they are, the comparisons between A and B , A and C as well as B and C . Compute the standard error of mean with formula (7.15) for equal sample sizes $n = 4$ and $S_{\bar{X}} = 0.51$. Then sort the means in ascending order, $\bar{X}_C = 9.3$, $\bar{X}_B = 11.3$, $\bar{X}_A = 16.0$. Given $a = 2, 3$ and $\alpha = 0.01$, refer to Table 7 of Appendix II, compute the critical values $q_{a,a,v} S_{\bar{X}}$.

For $a = 2$, $q_{0.01,2,9} = 4.60$, $q_{0.01,2,9} S_{\bar{X}} = 4.60 \times 0.51 = 2.35$.

For $a = 3$, $q_{0.01,3,9} = 5.43$, $q_{0.01,3,9} S_{\bar{X}} = 5.43 \times 0.51 = 2.77$.

To test $H_0 : \mu_B = \mu_C$, with $a = 2$, and the critical value $q_{0.01,2,9} S_{\bar{X}} = 2.35$, $\bar{X}_B - \bar{X}_C = 11.3 - 9.3 = 2.0 < 2.35$, $P > 0.01$, H_0 is not rejected.

To test $H_0 : \mu_A = \mu_B$, $\bar{X}_A - \bar{X}_B = 16.0 - 11.3 = 4.7 > 2.35$, $P < 0.01$, H_0 is rejected.

Anticoagulants	C	B	A
Means of ESR	9.3	11.3	16.0
Grouping	(C, B)		(A)

Fig. 7.10 The grouping of the effects of the three anticoagulants.

To test $H_0 : \mu_A = \mu_C$, with $a = 3$, and the critical value $q_{0.01,3,9}S_{\bar{X}} = 2.77$, $\bar{X}_A - \bar{X}_C = 16.0 - 9.3 = 6.7 > 2.77$, $P < 0.01$, H_0 is rejected.

All comparison results can simply be demonstrated in Fig. 7.10.

Figure 7.10 implies that the effects of the three anticoagulants to ESR can be divided into two classes, one is (C, B), and another is (A). In conclusion, there is no statistically significant difference between anticoagulant B and anticoagulant C, but anticoagulant A has different effect on ESR comparing with anticoagulant B and anticoagulant C.

7.1.5.3 Bonferroni test

Bonferroni pointed out that for each test, the test level is α' , if total m comparisons were performed, when H_0 is true, the cumulated probability for type I error $\leq m\alpha'$. This is the famous Bonferroni inequality. For example, after analysis of variance, the difference among the four samples has statistical significance. Then multiple comparisons between any two of the groups are needed. If there are three comparisons, $m = 3$ and $\alpha' = 0.05$, the probability of none of type I error for the three comparisons is $(1 - 0.05)^3 = 0.857$, and the cumulated probability for type I error is $1 - 0.857 = 0.143$. Hence, if the cumulated probability for type I error after multiple comparisons is limited by α , the above Bonferroni inequality $\alpha = m\alpha'$ can be used to ascertain the test level

$$\alpha' = \frac{\alpha}{m} \quad (7.18)$$

for each test.

$$t = \frac{|\bar{X}_i - \bar{X}_j|}{S_{\bar{X}_i - \bar{X}_j}} = \frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{MS_{\text{Error}} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}}, \quad v = v_{\text{error}}. \quad (7.19)$$

Basically, Bonferroni test is the adjustment of test level, so it is called Bonferroni adjustment method. The method is suitable for all comparison of two groups, including multiple comparisons of means and probabilities.

Table 7.8 Analysis table of t -test for multiple comparison of data in Example 7.2.

Comparison groups	Mean difference ($\bar{X}_i - \bar{X}_j$)	SE	t	P
A vs. B	4.75	0.726	6.543	<0.0167
A vs. C	6.75	0.726	9.298	<0.0167
B vs. C	2.00	0.726	2.755	0.067

Bonferroni test is the most conservative comparison method. When the time of comparisons (m) is less, the effect is better. However, if the time of comparisons (m) is more (for example, $m > 10$), the conclusion might be less conservative. Therefore, Šidák (1967) used

$$\alpha' = 1 - \sqrt[m]{1 - \alpha} \quad (7.20)$$

as the test level for each comparison.

Example 7.2 (Cont'd) Multiple comparison for the three anticoagulants.

Solution The H_0 was rejected in analysis of variance in Table 7.4, which indicated at least two groups had difference. (7.19) was used to perform t -test. According to Bonferroni test, the test level for each comparison is

$$\alpha' = \frac{\alpha}{m} = \frac{0.05}{3} = 0.0167$$

and the comparison results are listed in Table 7.8.

When $\alpha' = 0.0167$, the result of Bonferroni test showed statistical significance between A and B , A and C . However, no statistical significance was found between B and C .

7.2 Two-Way Analysis of Variance: Randomized Complete-Block Design

7.2.1 The randomized complete-block design

There are n blocks and each block contains G experimental units to receive G treatments randomly. The total number of observations is $N = nG$. The object of grouping the experimental units into n blocks before delivering treatments is to have the units in the same block as uniform as possible so

that the variability among different blocks will be greater than that within the blocks. The advantage of this design comparing with the completely random design is to reduce the effect of the variation among the experimental units if the difference among blocks was a main source of variation. The disadvantage is that all the sizes of different blocks (or say, the numbers of experimental units in different blocks) should be equal to the number of treatments, otherwise, the statistical analysis will be difficult. The principle of the randomized complete-block design has been used in paired *t*-test in comparison of two treatments with paired observations mentioned in Chap. 4, where each block contains only two units, that is, $G = 2$.

Example 7.4 12 mice have been grouped into four blocks according to their birth litters and each block has 3 mice. Randomly assign 3 kinds of feed to the 3 mice in each block.

Solution Number the 3 mice in each block in some convenient manner, say, assigning 1 to 3 to the mice according to the orders of their weights in each litter. The treatments are also numbered, for instance, feed *A*, feed *B* and feed *C* are represented by 1, 2 and 3 respectively. Then determine the numbers of treatments by the ranks of random numbers and repeat the procedures for each block. The results of randomization in this example are listed in Table 7.9.

7.2.2 Two-way analysis of variance

The observations after experiment, denoted as X_{ij} , can be listed in the format as Table 7.10. There X_{ij} means the reading in *i*th block, *j*th treatment; B_i is the sum of the readings in *i*th block; T_j is the sum of the readings in *j*th

Table 7.9 Random allocation of the randomized complete-block design.

Block	Unit No.			Random No. (Rank)			Unit No. (Treatment)		
1	1	2	3	28 (1)	65 (3)	62 (2)	1 (A)	2 (C)	3 (B)
2	1	2	3	79 (3)	21 (2)	05 (1)	1 (C)	2 (B)	3 (A)
3	1	2	3	81 (2)	51 (1)	94 (3)	1 (B)	2 (A)	3 (C)
4	1	2	3	19 (1)	90 (3)	76 (2)	1 (A)	2 (C)	3 (B)

Table 7.10 The observations of the randomized complete-block design.

Block	Treatment				Mean (block)
	1	2	...	k	
1	X_{11}	X_{12}	...	X_{1k}	$\bar{x}_{.1}$
2	X_{21}	X_{22}	...	X_{2k}	$\bar{x}_{.2}$
...
...
...
...
n	X_{n1}	X_{n2}	...	X_{nk}	$\bar{x}_{.n}$
Mean (treatment)	$\bar{x}_{.1}$	$\bar{x}_{.2}$...	$\bar{x}_{.k}$	\bar{x}

Table 7.11 The table of analysis of variance for randomized complete-block design.

Source	DF	SS	MS	F
Treatment	$k - 1$	SS_{B1}	$MS_{B1} = SS_{B1}/(k - 1)$	MS_{B1}/MS_E
Block	$n - 1$	SS_{B2}	$MS_{B2} = SS_{B2}/(n - 1)$	MS_{B2}/MS_E
Error	$(k - 1)(n - 1)$	SS_E	$MS_E = SS_E/(k - 1)(n - 1)$	
Total	$kn - 1$	SS_T		

treatment. Other calculations to form the table of analysis of variance are shown in Table 7.11.

(7.21)–(7.24) show the variance analyzing for two-way analysis of variance.

$$\begin{aligned}
 SS \text{ for total} &= \sum_{i=1}^n \sum_{j=1}^k (X_{ij} - \bar{X})^2 \\
 &= \text{sum of square for difference between experiment} \\
 &\quad \text{results and total mean,} \quad (7.21)
 \end{aligned}$$

$$\begin{aligned}
 SS \text{ for blocks} &= \sum_{i=1}^n k(\bar{X}_i - \bar{X})^2 \\
 &= \text{weighted sum of square for difference between each} \\
 &\quad \text{block mean and total mean} \quad (7.22)
 \end{aligned}$$

of which, k is the experiment units number represented by mean of each block.

$$SS \text{ for treatment groups} = \sum_{j=1}^k n(\bar{X}_{.j} - \bar{X})^2$$

= weighted sum of square for difference between
each treatment mean and total mean (7.23)

of which, n is the experiment units number represented by mean of each treatment group.

$$SS \text{ for error} = SS \text{ for total} - SS \text{ for blocks} - SS \text{ for treatment groups.} \quad (7.24)$$

Then, table of analysis of variance can be obtained. In Table 7.11, there are two F values which can be calculated for treatment and block respectively. Similarly, when $k = 2$, F_1 is equivalent to t value in paired t -test. In fact, the paired t is the positive square root of F_1 , that is, paired $t = \sqrt{F_1}$.

Example 7.5 The riboflavin concentration of three groups of broccoli leaves was measured under four conditions: A , B , C and D . The experiment results were listed in Table 7.12. Is there any statistical significance among the four conditions?

Solution There are three blocks for the three groups, $n = 3$, $k = 4$, $nk = 3 \times 4 = 12$. The descriptive statistical values of Table 7.12 are (7.21)–(7.24)

Table 7.12 The riboflavin concentration of broccoli leaves ($\mu\text{g/g}$).

Groups	Measurement conditions				Block means ($\bar{x}_{i.}$)
	A	B	C	D	
1	27.2	24.6	39.5	38.6	32.5
2	23.2	24.2	43.1	39.5	32.5
3	24.8	22.2	45.2	33.0	31.3
Mean ($\bar{x}_{.j}$)	25.1	23.7	42.6	37.0	32.1 (\bar{x})

for variance analyzing.

$$SS \text{ for total} = \sum_{i=1}^3 \sum_{j=1}^4 (X_{ij} - 32.1)^2 = 818.37.$$

$$\begin{aligned} SS \text{ for blocks} &= \sum_{i=1}^3 k(\bar{X}_{i.} - \bar{X})^2 \\ &= 4 \times (32.5 - 32.1)^2 + 4 \times (32.5 - 32.1)^2 \\ &\quad + 4 \times (31.3 - 32.1)^2 \\ &= 3.76. \end{aligned}$$

$$\begin{aligned} SS \text{ for treatment groups} &= \sum_{j=1}^k n(\bar{X}_{.j} - \bar{X})^2 \\ &= 3 \times (32.5 - 32.1)^2 + 3 \times (32.5 - 32.1)^2 \\ &\quad + 3 \times (31.3 - 32.1)^2 + 3 \times (31.3 - 32.1)^2 \\ &= 765.53. \end{aligned}$$

$$SS \text{ for error} = 818.37 - 3.76 - 765.53 = 49.08.$$

(Notes: The calculation results of SS for blocks and for treatment groups are just for demonstration. We only kept up to two decimal points, the results might different from that in Table 7.13 for rounding error.)

Then the table for two-way analysis of variance can be obtained (Table 7.13). The conclusion is that the difference among the four measurement conditions has statistical significance ($P < 0.01$), and difference among the three groups has no statistical significance.

Table 7.13 Table for two-way analysis of variance.

Source	<i>DF</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>P</i>
Measurement conditions	3	765.53	255.18	31.20	<0.01
Groups	2	3.76	1.88		
Error	6	49.08	8.18		
Total	11	818.37			

Residual analysis: using means of treatment groups as horizontal axis, the independence, normal distribution homogeneity of variances are shown in Figs. 7.11–7.13 respectively. It shows the experiment results accord with the basic assumption of two-way analysis of variance.

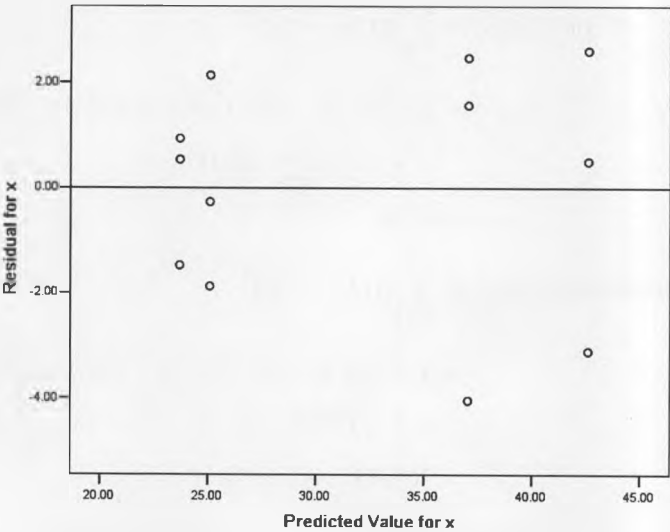


Fig. 7.11 Scatter plot for check of independence assumption for Table 7.12.

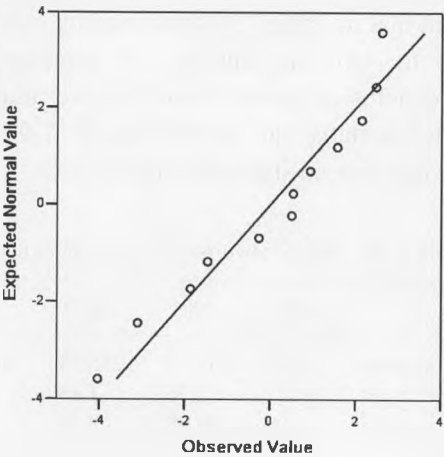


Fig. 7.12 Scatter plot for check of normal distribution for Table 7.12.

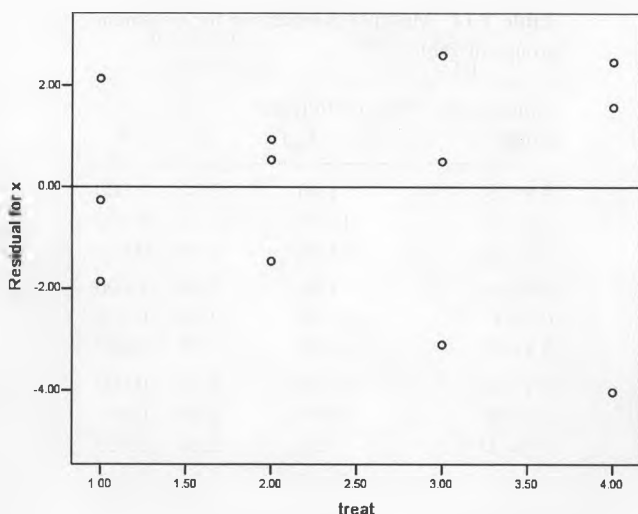


Fig. 7.13 Scatter plot for check of homogeneity of variances for Table 7.12.

The results for multiple comparison (Bonferroni test) of treatment groups (measurement conditions) are listed in Table 7.14 (from statistical software).

7.3 Three-Way Analysis of Variance: The Latin-Square Design

7.3.1 Design of Latin squares

Example 7.6 To compare the effects of three feeds to the weight of mice. There are three litters of mice and they have different weight initially.

Solution Refer to Fig. 7.14, in which the treatments are feeds denoted by A , B and C , and the mice are blocked by two systems, the rows represent litters the mice belonging to, and the columns represent the orders of weight in each litter.

In Fig. 7.14: $3 \times 3 = 9$ mice in total need to be observed; according to the order of weight, 3 mice in litter 1 are fed by A , B and C ; 3 mice in litter 2 are fed by B , C and A ; and 3 mice in litter 3 are fed by C , A and B .

Figure 7.14 is called 3×3 Latin square, where each row as well as each column has A , B and C appearing once and only once. In general, for any integer k one can have at least one $k \times k$ Latin square, in which each row

Table 7.14 Multiple comparison for treatment groups in Table 7.12.

Comparison groups	Mean difference $(\bar{X}_i - \bar{X}_j)$	SE	P
A vs. B	1.40	2.34	1.000
A vs. C	-17.53	2.34	0.002
A vs. D	-11.97	2.34	0.013
B vs. A	-1.40	2.34	1.000
B vs. C	-18.93	2.34	0.001
B vs. D	-13.37	2.34	0.007
C vs. A	17.53	2.34	0.002
C vs. B	18.93	2.34	0.001
C vs. D	5.57	2.34	0.327
D vs. A	11.97	2.34	0.013
D vs. B	13.37	2.34	0.007
D vs. C	-5.57	2.34	0.327

The order of weights

		I	II	III
	1	A	B	C
Litter	2	B	C	A
	3	C	A	B

Fig. 7.14 An example of 3×3 Latin square.

as well as each column has different Latin letters appearing once and only once. The design based on Latin square is a straightforward generalization of the randomized block design.

Example 7.7 Five subcutaneous injections with different dosages are injected into the bodies of rabbits on test their effects to the blister on the rabbits' skin. Make an experimental design with a Latin-square and test the differences among drugs.

Solution Take the positions of injection as experimental units, which can be simultaneously blocked in two ways, rabbits and positions in each rabbit. Selecting a basic 5×5 Latin-square from Table 7 of Appendix II, take five subcutaneous injections as treatments denoted by A, B, C, D and E, the

row blocks represent the rabbits labeled with 1, 2, 3, 4 and 5, the column blocks as the positions of injection labeled with I, II, III, IV and V. The procedures of random permutation of rows and columns of the basic 5×5 Latin-square table are demonstrated as follows.

7.3.1.1 Randomly permute the rows

Get four two-digit random numbers, say 66, 05, 32, 88, and rank them, $R = 3, 1, 2, 4$; exchange the third row with the first row, and exchange the second row with the fourth row.

A	B	C	D	E		C	D	E	A	B		C	D	E	A	B
B	C	D	C	A		B	C	D	E	A		D	E	A	B	C
C	D	E	A	B	$3 \leftrightarrow 1$	A	B	C	D	E	$2 \leftrightarrow 4$	A	B	C	D	E
D	E	A	B	C		D	E	A	B	C		B	C	D	E	A
E	A	B	C	D		E	A	B	C	D		E	A	B	C	D

7.3.1.2 Randomly permute columns

Get another four random numbers and their ranks, say 53, 85, 39, 06, $R = 3, 4, 2, 1$; exchange between columns 3 and 4 and between columns 2 and 1.

C	D	E	A	B		C	D	A	E	B		D	C	A	E	B
D	E	A	B	C		D	E	B	A	C		E	D	B	A	C
A	B	C	D	E	$3 \leftrightarrow 4$	A	B	D	C	E	$2 \leftrightarrow 1$	B	A	D	C	E
B	C	D	E	A		B	C	E	D	A		C	B	E	D	A
E	A	B	C	D		E	A	C	B	D		A	E	C	B	D

The final experimental design is listed in Table 7.15.

Another way of randomization in the Latin-square design is just randomly assign t treatments into t Latin letters in the basic Latin square when t is relative large, say $t \geq 6$.

7.3.2 Analysis of variance

The observations of Latin-square design could be listed as Table 7.16. Descriptive statistics $\bar{x}_{i.}$ is the mean of X_{ij} in row i , and $\bar{x}_{.j}$ is the mean of X_{ij} in column j , $\bar{x}_{.k}$ is the mean of X_{ij} under the treatment k .

Table 7.15 5×5 Latin-square design.

Number of rabbit	Number of injection position				
	I	II	III	IV	V
1	D	C	A	E	D
2	E	D	B	A	C
3	B	A	D	C	E
4	C	B	E	D	A
5	A	E	C	B	D

*Five subcutaneous injections denoted as A, B, C, D and E

Table 7.16 The experimental results of Latin-square design.

Row	Column				Mean (row)
	1	2	...	t	
1	X_{11}	X_{12}	...	X_{1t}	$\bar{x}_{1.}$
2	X_{21}	X_{22}	...	X_{2t}	$\bar{x}_{2.}$
...
T	X_{t1}	X_{t2}	...	X_{tt}	$\bar{x}_{t.}$
Mean (column)	$\bar{x}_{.1}$	$\bar{x}_{.2}$...	$\bar{x}_{.t}$	\bar{x}
Mean (treatment)	A	B	...		
Total	\bar{x}_1	\bar{x}_2	...	\bar{x}_t	

Equations (7.25)–(7.29) give the variance analyses for design of Latin squares.

$$\begin{aligned}
 SS \text{ for total} &= \sum_{i=1}^t \sum_{j=1}^t (X_{ij} - \bar{X})^2 \\
 &= \text{sum of square for difference between experiment results} \\
 &\quad \text{and total mean}
 \end{aligned}
 \tag{7.25}$$

$$\begin{aligned}
 SS \text{ for row} &= \sum_{i=1}^t t(\bar{X}_i - \bar{X})^2 \\
 &= \text{weighted sum of square for difference between each row} \\
 &\quad \text{mean and total mean}
 \end{aligned}
 \tag{7.26}$$

of which, t is the experiment units number represented by the mean of each row.

$$\begin{aligned}
 SS \text{ for column} &= \sum_{j=1}^k t(\bar{X}_{.j} - \bar{X})^2 \\
 &= \text{weighted sum of square for difference between each} \\
 &\quad \text{column mean and total mean} \quad (7.27)
 \end{aligned}$$

of which, t is the experiment units number represented by the mean of each column.

$$\begin{aligned}
 SS \text{ for treatment} &= \sum_{k=1}^t t(\bar{X}_k - \bar{X})^2 \\
 &= \text{weighted sum of square for difference between each} \\
 &\quad \text{treatment mean and total mean} \quad (7.28)
 \end{aligned}$$

of which, t is the experiment units number represented by the mean of each treatment group.

$$\begin{aligned}
 SS \text{ for error} &= SS \text{ for total} - SS \text{ for row} - SS \text{ for column} \\
 &\quad - SS \text{ for treatment.} \quad (7.29)
 \end{aligned}$$

Then, table of three-way analysis of variance can be obtained. In Table 7.17, there are three F values, which represent mean differences among treatments, rows and columns, respectively.

Example 7.8 In order to compare the effect of the seven kinds of drugs, seven diastasis intestine tracts of animals were used to receive the treatments of seven drugs for each intestine tract. The measurements were scores of

Table 7.17 The ANOVA table for Latin-square design.

Source	DF	SS	MS	F
Treatment	$t - 1$	SS_{B1}	$MS_{B1} = SS_{B1}/(t - 1)$	MS_{B1}/MS_E
Row	$t - 1$	SS_{B2}	$MS_{B2} = SS_{B2}/(t - 1)$	MS_{B2}/MS_E
Column	$t - 1$	SS_{B3}	$MS_{B3} = SS_{B3}/(t - 1)$	MS_{B3}/MS_E
Error	$(t - 1)(t - 2)$	SS_E	$MS_E = SS_E/(t - 1)(t - 2)$	
Total	$t^2 - 1$	SS_T		

Table 7.18 The design and the measured scores of response intensity.

Intestine tract	Order of treatments							Mean (row) ($\bar{x}_{i.}$)
	1	2	3	4	5	6	7	
1	A21	B19	C0	D0	E5	F5	G2	7.4
2	B25	E4	A3	G0	F1	D2	C0	5.0
3	C0	F7	G0	B11	D7	A6	E4	5.0
4	D10	G4	E7	F7	C0	B17	A7	7.4
5	E6	D0	B9	C0	A1	G4	F5	3.6
6	F7	C0	D10	A11	G3	E6	B15	7.4
7	G3	A6	F3	E12	B26	C0	D6	8.0
Mean (column)	10.3	5.7	4.6	5.9	6.1	5.7	5.6	
Mean (treatment) ($\bar{x}_{k.}$)	A	B	C	D	E	F	G	6.3 (\bar{x})
	55	122	0	35	44	35	16	

response intensity. Randomly assign seven drugs to be treatments denoted as A, B, C, D, E, F, G , and took intestine tracts as row block, the order of receiving the treatments as column block. A 7×7 Latin-square design was applied and the results were showed in Table 7.18. Judge whether there is any difference for the average scores of response intensity among the drugs.

Solution According to (7.25)–(7.29), we have

$$\text{Total SS} = \sum_{i=1}^7 \sum_{j=1}^7 (X_{ij} - 6.3)^2 = 2019.55.$$

$$\begin{aligned} \text{Row SS} &= \sum_{i=1}^7 7 \times (\bar{X}_{i.} - \bar{X})^2 \\ &= 7 \times (7.3 - 6.4)^2 + 7 \times (5.0 - 6.4)^2 + 7 \times (5.0 - 6.4)^2 \\ &\quad + 7 \times (7.4 - 6.4)^2 + 7 \times (3.6 - 6.4)^2 \\ &\quad + 7 \times (7.4 - 6.4)^2 + 7 \times (8.0 - 6.4)^2 \\ &= 122.69. \end{aligned}$$

$$\begin{aligned} \text{Column SS} &= \sum_{j=1}^7 7 \times (\bar{X}_{.j} - \bar{X})^2 \\ &= 7 \times (10.3 - 6.4)^2 + 7 \times (5.7 - 6.4)^2 + 7 \times (4.6 - 6.4)^2 \end{aligned}$$

$$\begin{aligned}
&+ 7 \times (5.9 - 6.4)^2 + 7 \times (6.1 - 6.4)^2 \\
&+ 7 \times (5.7 - 6.4)^2 + 7 \times (5.6 - 6.4)^2 \\
&= 142.12.
\end{aligned}$$

$$\begin{aligned}
\text{Treatment } SS &= \sum_{k=1}^7 7 \times (\bar{X}_k - \bar{X})^2 \\
&= 7 \times (7.9 - 6.4)^2 + 7 \times (17.4 - 6.4)^2 + 7 \times (0 - 6.4)^2 \\
&\quad + 7 \times (5.0 - 6.4)^2 + 7 \times (6.3 - 6.4)^2 \\
&\quad + 7 \times (5.0 - 6.4)^2 + 7 \times (2.3 - 6.4)^2 \\
&= 1298.12.
\end{aligned}$$

$$\text{Error } SS = 2019.55 - 129.69 - 142.12 - 1298.12 = 456.62.$$

(Notes: We only keep up to two decimal places for the results of SS for row, column and treatment groups, the results might be different from that of statistical software for rounding error.)

Then the table of analysis of variance can be obtained (Table 7.19). The conclusion is that the difference among the drugs has statistical significance ($P < 0.01$), and difference among the samples and orders has no statistical significance.

The Latin-square design can greatly reduce the experimental units, and especially suitable for animal or laboratory experiment. The disadvantage of this experimental design is that most experiments, for instance in clinical trials, could not meet the condition of the equal numbers of treatments, row blocks and column blocks.

Table 7.19 The ANOVA table for 7×7 Latin-square.

Source	<i>DF</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>P</i>
Among drugs	6	1298.12	216.35	14.21	<0.01
Among samples	6	122.96	20.45	1.34	>0.05
Among orders	6	142.12	23.69	1.56	>0.05
Error	30	456.62	15.22		
Total	48	2019.55			

Residual analysis: using means of treatment groups as horizontal axis, independence, normal distribution and homogeneity of variances can be shown in Figs. 7.15–7.17, respectively. Figure 7.15 shows sector trend, Fig. 7.16 shows some scatters departure from the line. The residual analysis

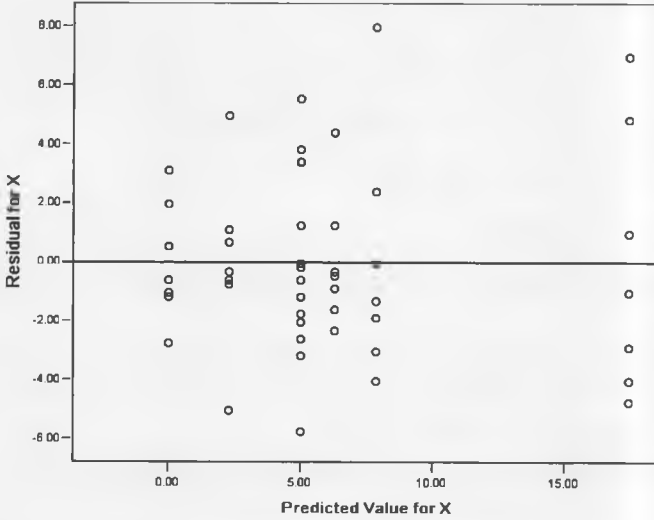


Fig. 7.15 Scatter plot for check of independence assumption for Table 7.18.

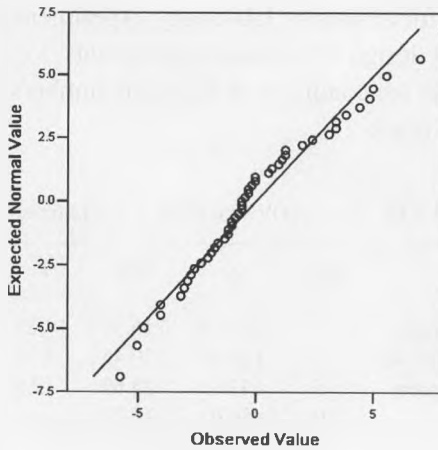


Fig. 7.16 Scatter plot for check of normal distribution for Table 7.18.

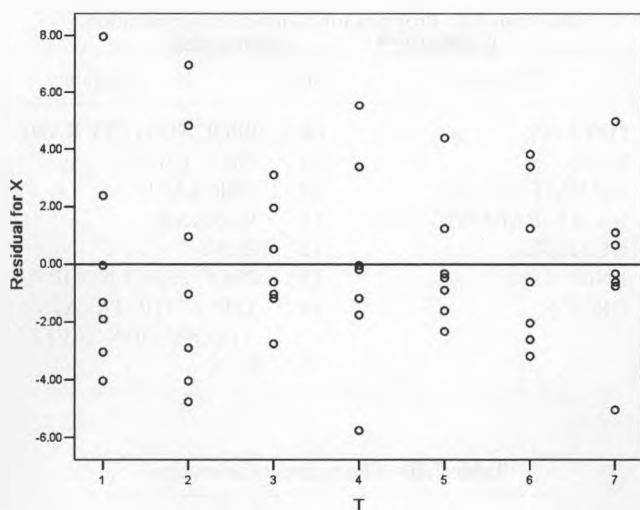


Fig. 7.17 Scatter plot for check of homogeneity of variances for Table 7.18.

Treatment (drugs)	C	G	D	F	E	A	B
Mean of intensity score (I class)	0	2.29	5.0	5.0			
Mean of intensity score (II class)		2.29	5.0	5.0	6.29	7.86	
Mean of intensity score (III class)							17.43

Fig. 7.18 Categories for different intensity of drugs effect.

after transformation can be used to improve the independence and normality (see Sec. 7.5 in this chapter).

The results of multiple comparisons (SNK test) for treatment groups can be found in Fig. 7.18 (from statistical software). There are three categories for the seven drugs: no effect or low effect (class I), middle effect (class II), and strong effect (class III).

7.4 Computerized Experiments

Experiment 7.1 Randomization with random numbers In Program 7.1, N is the number of objects in an experiment, and given 50 in current experiment. Table 7.20 shows the output of SAS program. Lines 03–06 in Program 7.1 create random numbers from a uniform distribution, noted

Program 7.1 Program for complete randomization.

Line	Program	Line	Program
01	DATA DS;	08	PROC SORT; BY RAND; RUN;
02	N=50;	09	PROC RANK;
03	DO UNIT=1 TO N;	10	VAR RAND;
04	RAND=RANUNI(12345);	11	RANKS R;
05	OUTPUT;	12	RUN;
06	END;	13	PROC PRINT NOOBS;
07	DROP N;	14	TITLE 'THE TREATMENT ALLOCATION TABLE';
		15	RUN;

Table 7.20 The treatment allocation.

UNIT	RAND	R	UNIT	RAND	R
32	0.02057	1	⋮	⋮	⋮
40	0.05069	2	10	0.76408	40
⋮	⋮	⋮	41	0.76474	41
23	0.35052	20	42	0.78983	42
35	0.35687	21	⋮	⋮	⋮
1	0.36292	22	34	0.98701	50

as 'RAND'; the number 12345 in line 04 is called seed, which can be transferred to any five-digital number. Line 08 is a procedure to create ranks of random numbers; lines 09–12 present ranked serial numbers named 'R'. Lines 13 and 14 control the output. From the output in Table 7.20, the experimenters can write their own allocation table, for example, the units with numbers of 32, 40, . . . , 23 and ranks 1–20 are allocated to group 1, the units with numbers of 35, 1, . . . , 10 and ranks 21–40 to group 2, the units with numbers of 41, 42, . . . , 34 and ranks 41–50 to group 3.

Experiment 7.2 One-way analysis of variance (completely randomized design) Program 7.2 is a program to analyze data with completely randomized design in Example 7.3. Line 03 in Program 7.2 could be $Y1=X$ (no transformation) or $Y1=\text{SQRT}(X+1)$. Lines 09 and 10 control the

Program 7.2 Program for analyzing Example 7.3.

Line	Program	Line	Program
01	DATA FEVER;	09	PROC MEANS;VAR X Y1;
02	INPUT TREAT \$ X @@;	10	CLASS TREAT; RUN;
03	Y1=SQRT(X);	11	PROC ANOVA;
04	CARDS;	12	CLASS TREAT;
05	A 0 A 2 A 0 A 0 A 5 A 9	13	MODEL Y1=TREAT;
06	B 32 B 13 B 6 B 7 B 10 B 2	14	MEANS TREAT/SNK;
07	C 0 C 11 C 15 C 11 C 3 C 1	15	RUN;
08	;		

printing mean and standard deviation. Lines 11–13 control the printing of ANOVA table. Line 14 indicates the method of multiple comparisons, where SNK could be replaced by LSD (the least significant test) if necessary.

Experiment 7.3 Residual analysis Program 7.3 is a program for residual analysis after $y_1 = \sqrt{x}$ transformation for Example 7.3. Lines 09–13 use the GLM to fit complete randomized analysis of variance model and create GMEANS and residual ERR. Lines 14–16 keep the basic data set of residual analysis which have three analysis variables: treatment group

Program 7.3 Program for residual analysis for Example 7.3.

Line	Program	Line	Program
01	DATA FEVER;	14	DATA ResidualPlot;
02	INPUT TREAT \$ X @@;	15	SET Residual (KEEP=TREAT
03	Y1=SQRT(X);		GMEANS ERR);
04	CARDS;	16	RUN;
05	A 0 A 2 A 0 A 0 A 5 A 9	17	PROC PLOT DATA=ResidualPlot;
06	B 32 B 13 B 6 B 7 B 10 B 2	18	PLOT ERR*GMEANS='*' /
07	C 0 C 11 C 15 C 11 C 3 C 1		ERR*TREAT='*' / vref = 0 vpos=25
08	;		hpos=60 ;
09	PROC GLM;	19	RUN;
10	CLASS TREAT;	20	PROC CAPABILITY DATA
11	MODEL Y1=TREAT;	21	=ResidualPlot;
12	OUTPUT OUT=Residual		QQPLOT ERR/ NORMAL (mu=0
	predicted=GMEANS residual=ERR;		sigma=1 color=blue);
13	RUN;	22	RUN;

Program 7.4 Program for analyzing Example 7.7.

Line	Program	Line	Program
01	DATA LATIN;	13	5 4 C 0 5 5 A 1 5 6 G 4 5 7 F 5
02	INPUT R C TREAT \$ X @@;	14	6 1 F 7 6 2 C 0 6 3 D 1 0
03	CARDS;	15	6 4 A 1 1 6 5 G 3 6 6 E 6 6 7 B 1 5
04	1 1 A 2 1 1 2 B 1 9 1 3 C 0	16	7 1 G 3 7 2 A 6 7 3 F 3
05	1 4 D 0 1 5 E 5 1 6 F 5 1 7 G 2	17	7 4 E 1 2 7 5 B 2 6 7 6 C 0 7 7 D 6
06	2 1 B 2 5 2 2 E 4 2 3 A 3	18	;
07	2 4 G 0 2 5 F 1 2 6 D 2 2 7 C 0	19	PROC ANOVA;
08	3 1 C 0 3 2 F 7 3 3 G 0	20	CLASS R C TREAT;
09	3 4 B 1 1 3 5 D 7 3 6 A 6 3 7 E 4	21	MODEL X = R C TREAT;
10	4 1 D 1 0 4 2 G 4 4 3 E 7	22	MEANS TREAT/LSD;
11	4 4 F 7 4 5 C 0 4 6 B 1 7 4 7 A 7	23	RUN;
12	5 1 E 6 5 2 D 0 5 3 B 9		

TREAT, Fitting value GMEANS, residual ERR. Lines 17–20 are used to show residual plot for check of independence (ERR*GMEANS), and for check of homogeneity of variance (ERR*TREAT). Lines 21–23 are used to make Q–Q plot for check of normality (QQPLOT).

Experiment 7.4 Three-way analysis of variance (Latin-square design)

Experiment 7.5 Two-way analysis of variance (the randomized complete block design) Line 17 in Program 7.5 can be rewritten as MODEL X=TREAT without BLOCK. Readers can compare the difference in the outputs.

Program 7.5 Program for analyzing Example 7.5.

Line	Program	Line	Program
01	DATA RBK;	11	PROC MEANS;
02	INPUT BLOCK TREAT \$ X @@;	12	VAR X;
03	CARDS;	13	CLASS TREAT;
04	1 A 27.2 1 B 24.6	14	RUN;
05	04 1 C 39.5 1 D 38.6	15	PROC ANOVA;
06	2 A 23.2 2 B 24.207	16	CLASS BLOCK TREAT;
07	2 C 43.1 2 D 39.5	17	MODEL X=BLOCK TREAT;
08	3 A 24.8 3 B 22.2	18	MEANS TREAT/SNK;
09	3 C 45.2 3 D 33.0	19	RUN;
10	;		

7.5 Practice and Experiments

1. To compare the difference of G population means, why is it that sometimes we use complete randomized design and sometimes we use randomized complete-block design or Latin-square design? When to use SAS to analyze the result of experiment? What is the difference among the three designs in data formats, model selections and the methods of multiple comparisons?
2. To randomly assign 28 rabbits into four treatment groups A , B , C and D , and each group has equal number of samples, try to write a grouping process and give a result of grouping (with reference to Program 7.1 for complete randomized design given $N = 28$).
3. To study the influence of four feeds A , B , C and D to the weight of mice which can be grouped by 8 litters with 4 mice, which experiment design should be chosen? Try to write the process of randomization, grouping result and the source of variations in ANOVA table.
4. To compare the curative effect of a control, two Chinese traditional medicines and two Western medicines, they were applied to treat five infectious skin areas of each rabbit, what experiment design should be chosen? Try to write randomization procedure, allocation table and the source of variations in ANOVA table.
5. To perform the arc-sine of square root transformation for experiment results of Example 7.8, use residual analysis to compare the independence, normality and homogeneity of variance before and after the transformation.
6. Test the effects of estrogen with three dosages on the weights of uterus of female rats according to the observations in Table 7.21.
7. To compare the skin herpes size (mm^2) of the rabbit after injecting six types of drugs A , B , C , D , E , F . Six rabbits were selected and six different positions of the rabbits were injected in the study according to a Latin-square design. The experimental design and results are listed in Tables 7.22 and 7.23, respectively analyze the data.

Table 7.21 The weights of uterus of female rats (g).

Rats	Injection dosage of estrogen ($\mu\text{g}/100\text{g}$)		
	0.2	0.4	0.8
A	106	116	145
B	42	68	115
C	70	111	133
D	42	63	87

Table 7.22 Random allocation result of Latin-square design.

Injection position (row)	Rabbit (column)					
	1	2	3	4	5	6
1	A	B	E	F	C	D
2	B	A	D	C	F	E
3	C	E	F	B	D	A
4	E	F	C	D	A	B
5	D	C	B	A	E	F
6	F	D	A	E	B	C

Table 7.23 The result of the skin herpes size (mm^2) of rabbit with the injection of six drugs.

Injection position (row i)	Rabbit (column j)						$\sum X_i$	\bar{X}_i
	1	2	3	4	5	6		
1	73	83	73	58	64	77	428	71.3
2	75	81	60	64	62	75	417	69.5
3	67	99	73	64	64	73	440	73.3
4	61	82	77	71	81	59	431	71.8
5	69	85	68	77	85	85	469	78.2
6	79	87	74	74	71	82	467	77.8
$\sum X_j$	424	517	425	408	427	451		
\bar{X}_j	71.7	86.2	70.8	68.0	71.2	75.2		
Drug (k)	D	E	C	A	B	F		
$\sum X_k$	428	467	439	459	420	439	$\bar{X} = 73.7$	
\bar{X}_k	71.3	77.8	73.2	76.5	70.0	73.2		

8. To perform residual analysis after square root or logarithm transformation for Example 7.8, and check if the independence and normality are improved.

(1st edn. Yongyong Xu, Yi Wan, Jiqian Fang; 2nd edn. Yongyong Xu, Yi Wan, Jiqian Fang)



Chapter 8

Nonparametric Test Based on Ranks

The statistical tests we have discussed up to now, with one exception, are classified as the parametric statistics. These tests have two features: we have the knowledge of distribution of the population from which the samples were drawn, providing the basis for the inference; and our interest was focused on testing a hypothesis about one or more population parameters. An example of a parametric statistical test is the Student- t test. In order to use this test, the sampled population or populations are to be at least approximately normally distributed. One exception is the chi-square test, as a test of goodness-of-fit and as a test of independence.

In practice, the conditions for Student- t test are not always satisfied. For example, sometimes we are not sure if the samples come from normal distributions, or sometimes the distribution of sampled population is unknown. Therefore, the inference methods which are not depending on the distribution of population and the tests which are not focused on the hypotheses about the parameters of the populations will be considered. Those types of statistical inference procedures are called nonparametric statistics. As we will learn, the procedures that we discuss in this chapter either are not concerned with population parameters or do not depend on the knowledge of the sampled population. Since we do not make any specific assumption about the sampled population, those kinds of statistical inference procedures are also called distribution-free statistics.

The above discussion implies the following advantages of nonparametric statistics:

- (1) They could be widely used for different kinds of data because they do not depend on the distribution of sampled population and the testing hypotheses are not a statement about the population parameter.

- (2) Nonparametric tests could be used when the form of the sampled population is unknown.
- (3) Nonparametric test procedures tend to be easier in computation and consequently more quickly applied than the parametric procedures.
- (4) Nonparametric tests could be applied for ordinal variables or continuous variable given with rank only.

Although the nonparametric tests retain a number of advantages, their disadvantages must also be recognized. The use of nonparametric test with data that can be handled with a parametric test may result in a waste of some information of the data, which means a reduction of testing power.

There are a variety of nonparametric procedures. In this chapter, we will only focus on the nonparametric tests based on ranks.

8.1 Wilcoxon's Signed Rank Test

8.1.1 *The test for small sample*

The Wilcoxon's signed rank test (1945) is applied for testing if a population median equal to certain value or zero as well as for testing the difference of pairwise designed data.

Suppose that there is a matched sample of random variables X and Y with m pairs of individuals. The i th pair of individuals have the measured values (x_i, y_i) and the difference $d_i = x_i - y_i$, $i = 1, \dots, m$. The median of $D = X - Y$ is denoted by $Md(D)$.

The hypotheses to be tested are

$$H_0: Md(D) = 0 \quad H_1: Md(D) \neq 0. \quad (8.1)$$

The following steps are usually followed:

- (1) Calculate the differences $d_i = x_i - y_i$ for $i = 1, 2, \dots, m$ and ignore all the pairs with zero difference, say, resulting in n pairs with nonzero difference.
- (2) Rank the absolute values of nonzero d_i s from the smallest to the largest so that each $|d_i|$ is ranked; if two or more of the $|d_i|$ s are equal (we say there is a tie), each of the tied value should share the average rank accordingly. For example, the three smallest $|d_i|$ s are equal, initially

Table 8.1 Decision rules of Wilcoxon signed rank test.

	Two-side test	One-side test (1)	One-side test (2)
Test hypotheses	$H_0: Md(D) = 0$ $H_1: Md(D) \neq 0$	$H_0: Md(D) = 0$ $H_1: Md(D) > 0$	$H_0: Md(D) = 0$ $H_1: Md(D) < 0$
Statistical decision:			
Small sample size, check the table	If $T^* \leq T_{\alpha/2}(n)$, H_0 is rejected	If $T_- \leq T_{\alpha}(n)$, H_0 is rejected	If $T_+ \leq T_{\alpha}(n)$, H_0 is rejected
Large sample size, normal approximation	If $ Z > Z_{\alpha/2}$, H_0 is rejected	If $ Z > Z_{\alpha}$, H_0 is rejected	If $ Z > Z_{\alpha}$, H_0 is rejected

they may get rank 1, 2 and 3, but due to tie, finally each of the three should share a rank of $(1 + 2 + 3)/3 = 2$.

(3) Assign the initial signs of d_i s to their ranks.

(4) Find the sum of the ranks with positive signs and denote by T_+ ; find the sum of the ranks with negative signs and denote by T_- . Let $T^* = \min(T_+, T_-)$.

(5) Given the value of α , find the critical value $T_{\alpha/2}$ of the statistic from Table 10 in Appendix II. If $T^* \leq T_{\alpha/2}(n)$, H_0 is rejected.

For one-side tests, one can decide according to the rules given in Table 8.1.

8.1.2 The test for large sample (normal approximation)

When $n > 50$, Table 10 in Appendix II cannot be used. Then we turn to the normal approximation. In fact, it can be proved that if H_0 is true, when n is large enough, the distribution of statistic T^* will close to a normal distribution with

$$\mu_T = \frac{n(n+1)}{4} \quad (8.2)$$

and

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}. \quad (8.3)$$

If there is no tie, the statistic

$$Z = \frac{|T^* - \mu_T| - 0.5}{\sigma_T} = \frac{|T^* - n(n+1)/4| - 0.5}{\sqrt{n(n+1)(2n+1)/24}} \quad (8.4)$$

Table 8.2 Intelligence scores of 12 pairs of twin brothers.

Pair No. i	Senior x_i	Junior y_i	Differences $d_i = y_i - x_i$	Ranks of $ d_i R_i$	Ranks with sign $\pm R_i$
1	86	88	2	3	3
2	71	77	6	7	7
3	77	76	-1	1.5	-1.5
4	68	64	-4	4	-4
5	91	96	5	5.5	5.5
6	72	72	0	—	—
7	77	65	-12	10	-10
8	91	90	-1	1.5	-1.5
9	70	65	-5	5.5	-5.5
10	71	80	9	9	9
11	88	81	-7	8	-8
12	87	72	-15	11	-11

will follow a standard normal distribution. “0.5” in Eq. (8.4) is just the correction for continuity. The decision rules are also listed in Table 8.1.

If there are ties, Lehmann (1975) adjusted formula may be used for the statistic

$$Z_c = \frac{|T^* - n(n + 1)/4| - 0.5}{\sqrt{[n(n + 1)(2n + 1) - 0.5 \sum (t_p^3 - t_p)]/24}}$$

(8.5)

where t_p is the number of individuals in the p th tied subgroup.

Example 8.1 In order to study the difference of intelligence of twin brothers, the intelligence scores of 12 pairs of twin brothers were measured. The results are listed in Table 8.2.

Solution (1) Calculate the differences $d_i = x_i - y_i, i = 1, \dots, 12$, and ignore the sixth pair with zero difference, resulting in $n = 11$ pairs with nonzero difference, listed in column 4 of Table 8.2.

(2) Rank the absolute values of nonzero d_i s from the smallest to the largest in column 5 of Table 8.2; the difference of sixth pair is zero and is ignored from ranking. The differences of the third and eighth pair are equal and the average rank $(1 + 2)/2 = 1.5$ is shared by them, similar situations for the fifth pair and ninth pair.

(3) Assign the initial signs of d_i s to their ranks. See column 6 of Table 8.2.

(4) Calculate the sum of positive ranks and negative ranks respectively.

$$T_+ = 3 + 7 + 5.5 + 9 = 24.5,$$

$$T_- = 1.5 + 4 + 10 + 1.5 + 5.5 + 8 + 11 = 41.5.$$

We may check the calculation by

$$T_+ + T_- = \frac{n(n+1)}{2}.$$

In this case, $T_+ + T_- = 66 = 11(11+1)/2$, the ranking and calculation are correct. Choose the test statistic $T^* = \min(T_+, T_-) = 24.5$.

(5) Given the value of $\alpha = 0.05$, find the critical value $T_{0.05/2} = 11$ from Table 10 in Appendix II. Since $T^* > T_{0.05/2}$, $P > 0.05$, H_0 is not rejected.

Since $n = 11$ in this example, not a large sample, normal approximation is not proper. However, in order to show the calculation procedure of normal approximation, we just demonstrate the calculation of Z statistic here.

In this example, there are two values with ties. By Eq. (8.5),

$$\begin{aligned} Z_c &= \frac{24.5 - (11)(12)/4 - 0.5}{\sqrt{[(11)(12)(23) - (0.5)\{(2^3 - 2) + (2^3 - 2)\}]/24}} \\ &= \frac{8.5 - 0.5}{\sqrt{(3036 - 6)/24}} = \frac{8}{11.2361} = 0.7120. \end{aligned}$$

If the ties are not adjusted, by (8.4), Z will equal to 0.7112. One can see that the difference between Z and Z_c is very small and the value of Z increases after adjustment.

By checking the table of standard normal distribution, $P = 0.4756$, H_0 is not rejected.

8.1.3 Construction of the table for critical values

In order to explain how to construct Table 10 in Appendix II, let us work out an example.

Suppose there are $n = 4$ pairs of observations and their differences are not zero nor equal so that the ranks of the differences are 1, 2, 3, 4. When H_0 is true and the signs of d_i s are totally random, there are $2^4 = 16$ combinations with the same probability $P = 1/16 = 0.0625$. The distribution of all possible combinations and their rank sums are listed in Table 8.3. From Table 8.3 one can see that if $n = 4$, there is no value corresponding to the

Table 8.3 All possible rank sums and the distribution of T^* when $n = 4$.

Positive rank	Negative rank	Positive rank sum T_+	Negative rank sum T_-	Statistic T^*	Probability
1, 2, 3, 4	—	10	0	0	0.0625
2, 3, 4	1	9	1	1	0.0625
1, 3, 4	2	8	2	2	0.0625
1, 2, 4	3	7	3	3	0.1250
3, 4	1, 2	7	3	3	
1, 2, 3	4	6	4	4	0.1250
2, 4	1, 3	6	4	4	
1, 4	2, 3	5	5	5	0.1250
2, 3	1, 4	5	5	5	
1, 3	2, 4	4	6	4	0.1250
4	1, 2, 3	4	6	4	
1, 2	3, 4	3	7	3	0.1250
3	1, 2, 4	3	7	3	
2	1, 3, 4	2	8	2	0.0625
1	2, 3, 4	1	9	1	0.0625
—	1, 2, 3, 4	0	10	0	0.0625

cumulative probabilities 0.05, 0.025, 0.01 and 0.005. This is the reason why there is no critical value for $n = 4$.

In fact, for any other values of $n > 4$, one can work out the distribution of the statistic T^* in the similar way and find the values of T^* corresponding to the cumulative probabilities 0.05, 0.025, 0.01 and 0.005.

8.2 Wilcoxon's Rank-Sum Test for Comparing the Locations of Two Distributions

8.2.1 The test for small sample

The t test has been introduced for testing the difference of two population means in Chap. 4. When the assumptions underlying this technique are not met, that is, when the sampled populations are not normally distributed with equal variance or when the data for analysis consist only of ranks, a

nonparametric alternative to the t test may be used to test the hypothesis of equal location parameters.

Assume that two independent samples with sample size n_1 and n_2 are drawn respectively from two populations and the shapes of their distributions are similar. The medians of the two distributions are denoted by Md_1 and Md_2 .

The hypotheses to be tested are

$$H_0: Md_1 = Md_2 \quad H_1: Md_1 \neq Md_2.$$

The steps usually followed are:

- (1) Combine the two samples and rank all observations from the smallest to the largest while keeping track of sample to which each observation belongs. If there is tie, the average rank is shared by each.
- (2) Calculate the sum of the ranks for the two samples, denoted by R_1 and R_2 respectively.

If H_0 is true, R_1 and R_2 ought to be roughly proportional to their sample sizes; otherwise, they are not. However, in any case,

$$R_1 + R_2 = \frac{n(n+1)}{2}.$$

- (3) Determine the statistic T^* . When $n_1 < n_2$, let $T^* = R_1$; and when $n_1 = n_2$, let $T^* = \min(R_1, R_2)$.
- (4) Determine the critical values. Given the value of α , one can find the critical values T_α^L and T_α^U from Table 11 in Appendix II by $n_1 (< n_2)$ and $n_2 - n_1$.
- (5) The decision rules are: If $T^* \leq T_\alpha^L$ or $T^* \geq T_\alpha^U$, $P \leq \alpha$, H_0 is rejected; otherwise, if $T_\alpha^L < T^* < T_\alpha^U$, $P > \alpha$, H_0 is not rejected.

8.2.2 The test for large sample (normal approximation)

It can be proved that when sample size is big enough, the distribution of the statistic T^* closes to a normal distribution with

$$\mu_{T^*} = \frac{n_1(n+1)}{2}, \quad (8.6)$$

$$\sigma_{T^*}^2 = \frac{n_2 n_2 (n+1)}{12} \left[1 - \frac{\sum (t_p^3 - t_p)}{n^3 - n} \right]. \quad (8.7)$$

Table 8.4 Survival times of cats and rabbits without oxygen (min).

Cats		Rabbits	
Survival time	Rank	Survival time	Rank
25	9.5	15	1.5
34	15	15	1.5
44	17	16	3
46	18.5	17	4
46	18.5	19	5
		21	6.5
		21	6.5
		23	8
		25	9.5
		27	11
		28	12.5
		28	12.5
		30	14
		35	16
$n_1 = 5$	$R_1 = 78.5$	$n_2 = 14$	$R_2 = 111.5$

Here the part within brackets is the adjustment for ties.

Therefore, when H_0 is true, the test statistic

$$Z = \frac{|T^* - n_1(n+1)/2| - 0.5}{\sqrt{\sigma_{T^*}^2}} \quad (8.8)$$

will follow a standard normal distribution. Same as before, “0.5” plays the role of correction for continuity. Given the value of α , after checking the table of standard normal distribution, if the P -value corresponding to the value of Z is less than α , H_0 can be rejected.

Example 8.2 Without oxygen, the survival time (minute) of four cats and 14 rabbits are listed in Table 8.4. Now we try to compare the difference of survival times of cats and rabbits in the environment without oxygen.

Solution (1) Combine the two samples and rank all observations from smallest to largest in columns 2 and 4. There are two individuals taking the same value 15, of which the corresponding ranks are 1 and 2 so that the average rank $(1 + 2)/2 = 1.5$ is shared by each of them; again, there are two individuals taking the same value 25, of which the corresponding ranks

are 9 and 10 so that the average rank $(9 + 10)/2 = 9.5$ is shared by each of them.

(2) Calculate the sums of ranks for the two samples, denoted by R_1 and R_2 respectively.

$$R_1 = 9.5 + 15 + 17 + 18.5 + 18.5 = 78.5,$$

$$R_2 = 1.5 + 1.5 + \dots + 14 + 16 = 111.5,$$

$$R_1 + R_2 = 190 = \frac{19(19 + 1)}{2}.$$

It confirms the calculation.

(3) Determine the statistic T^* . Since $n_1 = 5 < n_2 = 14$, let $T^* = R_1 = 78.5$.

(4) Determine the critical values. From Table 11 in Appendix II, by $n_1 = 5$ and $n_2 - n_1 = 9$ we have $T_{0.01}^L = 22$, $T_{0.01}^U = 78$.

(5) Since $T^* = 78.5 > 78 = T_{0.01}^U$, $P < 0.01$, H_0 is rejected. It concludes that the survival times of cats and rabbits in the environment without oxygen might be different.

In this example, n_1 and n_2 are small, normal approximation is not suitable. However, in order to demonstrate the basic steps of normal approximation, let us work out the calculation as follows:

$$T^* = 78.5 \quad \mu_{T^*} = \frac{n_1(n+1)}{2} = \frac{5(19+1)}{2} = 50,$$

$$T^* - \mu_{T^*} = 78.5 - 50 = 28.5,$$

$$\begin{aligned} \sigma_{T^*}^2 &= \frac{n_2 n_2 (n+1)}{12} \left[1 - \frac{\sum (t_k^3 - t_k)}{n^3 - n} \right] \\ &= \frac{(5)(14)(19+1)}{12} \left[1 - \frac{(2^3 - 2) \times 5}{19^3 - 19} \right] \\ &= \frac{1400}{12} [0.995614035] = 116.1549708 \\ Z &= \frac{28.5 - 0.5}{\sqrt{116.1549708}} = 2.5980. \end{aligned}$$

Since $Z = 2.598 > Z_{0.01} = 2.58$, $P < 0.01$, H_0 is rejected.

Table 8.5 20 possible combinations and their rank sums in the first group when $n_1 = n_2 = 3$.

Rank	1, 2, 3	1, 2, 4	1, 2, 5	1, 2, 6	1, 3, 4	1, 3, 5	1, 3, 6	1, 4, 5	1, 4, 6	1, 5, 6
R_1	6	7	8	9	8	9	10	10	11	12
Rank	2, 3, 4	2, 3, 5	2, 3, 6	2, 4, 5	2, 4, 6	2, 5, 6	3, 4, 5	3, 4, 6	3, 5, 6	4, 5, 6
R_1	9	10	11	11	12	13	12	13	14	15

Table 8.6 The distribution of the rank sums in the first group when $n_1 = n_2 = 3$.

R_1	6	7	8	9	10	11	12	13	14	15
$P(R_1)$	0.05	0.05	0.10	0.15	0.15	0.15	0.10	0.10	0.05	0.05

8.2.3 The construction of the table for critical value

In order to explain how to construct Table 11 in Appendix II, let us also work out an example.

Suppose $n_1 = 3$ and $n_2 = 3$, when H_0 is true, the data can be pooled. For the data in the first group, there are $\binom{6}{3} = 20$ possible combinations of the ranks 1, 2, 3, 4, 5 and 6. The combinations and their rank sums are listed in Table 8.5.

The probability of each possible combination is equal to $1/20$ so that the probability distribution of the rank sums can be summarized as Table 8.6.

From this distribution, we may find the critical value $T_{0.05}^* = 6$, $T_{0.05}^U = 15$. These are exactly equal to those listed in Table 11 in Appendix II corresponding to $n_1 = 3$ and $n_2 - n_1 = 0$.

8.2.4 The comparison of ranked data

The Wilcoxon's rank-sum test for two independent samples is also useful for determining whether the values of one sample are higher than that of another sample even when the response variable was ordinal.

Example 8.3 In order to evaluate whether the improved-vaccine can enhance immunity of body, 200 volunteers were recruited and randomly divided into two groups, of which 100 subjects in group 1 was immunized by traditional vaccine (TV) and 100 subjects in group 2 received the improved-vaccine (IV). The antibody levels were detected one month after

Table 8.7 The comparison of antibody levels of the two groups.

Antibody level	TV (1)	IV (2)	Total (1) + (2) (3)	Range of ranks (4)	Average rank (5)	Sum of ranks	
						Pre (1) \times (5)	Post (2) \times (5)
—	18	3	21	1–21	11	198	33
\pm	20	10	30	22–51	36.5	730	365
+	46	14	60	52–111	81.5	3749	1141
++	11	50	61	112–172	142	1562	7100
+++	5	23	28	173–200	186.5	932.5	4289.5
Total	100	100	200	—	—	7171.5	12928.5

the final injection, and the result was showed in Table 8.7. Now the question is whether the difference of the antibody levels in the two groups is of statistical significance.

Solution (1) The test hypotheses are

H_0 : The medians of antibody level in the two populations are equal.

H_1 : The medians of antibody level in the two populations are different.

$$\alpha = 0.05.$$

(2) Compute the ranks for the data. First gather all of them with the same antibody level in two groups, as shown in Table 8.7. There are 21 with antibody level “—” who have a rank range of 1–21 and are assigned an average rank of $(1 + 21)/2 = 11$. There are 30 for the two groups combined with the antibody level “ \pm ”. The rank range for this group is from $(1 + 21)$ to $(30 + 21) = 22$ to 51. Thus all people in this group are assigned the average rank $= (22 + 51)/2 = 36.5$, and similarly for the other groups. By doing so, we can get the average ranks for the rest of antibody levels, 81.5, 142.5, 186.5.

(3) Calculate the sums of ranks. The rank sum for each group is equal to the sum of the product of the combined sample size and average rank of the same antibody level. For group 1,

$$\begin{aligned}
 R_1 &= (18 \times 11) + (20 \times 36.5) + (46 \times 81.5) \\
 &\quad + (11 \times 142) + (5 \times 186.5) \\
 &= 198 + 730 + 3749 + 1562 + 932.5 = 7171.5.
 \end{aligned}$$

For group 2,

$$\begin{aligned} R_2 &= (3 \times 11) + (10 \times 36.5) + (14 \times 81.5) \\ &\quad + (50 \times 142) + (23 \times 186.5) \\ &= 33 + 365 + 1141 + 7100 + 4289.5 = 12928.5. \end{aligned}$$

(4) Determine the critical values. In this case, $n_1 = 100$, $n_2 = 200$, the sample size is large enough, and normal approximation can be used here. According to Eq. (8.8), $Z = 7.264$, $P < 0.001$ (two-side), H_0 is rejected. And as $\bar{R}_1 = 7171.5/100 = 71.715 < \bar{R}_2 = 12928.5/100 = 129.285$, we can conclude that the improved-vaccine is better than the traditional one.

8.3 Hypothesis Testing for the Locations of More Than Two Populations

Here we will introduce the Kruskal–Wallis test for the data from a completely random design and the Friedman test for the data from a randomized block design.

8.3.1 *Kruskal–Wallis test for the data from a completely random design*

The one-way analysis of variance may be used to test the null hypothesis that several population means are equal when the variables follow normal distributions with equal variances. If the assumptions underlying the ANOVA are not met, or if the data for analysis consist only of ranks, a nonparametric alternative to the one-way analysis of variance may be used to test the hypothesis of equal location parameters. The one-way analysis of variance by ranks proposed by Kruskal–Wallis is the best known of these procedures.

Suppose that there are k independent samples drawn from k populations, and the sample sizes n_1, n_2, \dots, n_k may not be equal. The total sample size is n . The data may be listed as the format in Table 8.8, where x_{ij} refers to j th observation in sample i . We assume that the shapes of k distributions are similar and then test the hypotheses

$$H_0: Md_1 = Md_2 = \dots = Md_k$$

$$H_1: Md_1, Md_2, \dots, Md_k \text{ are not all equal.}$$

Table 8.8 The data format for completely random designed samples.

Number of groups			
1	2	...	k
x_{11}	x_{21}	...	x_{k1}
x_{12}	x_{22}	...	x_{k2}
\vdots	\vdots	\vdots	\vdots
x_{1n_1}	x_{2n_2}	...	x_{kn_k}

The main steps are:

- (1) Combine the k samples and rank all observations from smallest to largest while keeping track of sample to which each observation belongs. If there is tie, the average rank is shared by each.
- (2) Calculate the rank sums for the k samples, denoted by R_i , $i = 1, 2, \dots, k$ respectively.

$$R_i = \sum_{j=1}^{n_i} r_{ij} \quad i = 1, 2, \dots, k.$$

- (3) Calculate the test statistic.

When H_0 is true, the expectation and variance of the rank sum in sample i are

$$\mu_{R_i} = \frac{n_i(n+1)}{2}, \quad (8.9)$$

$$\sigma_{R_i}^2 = \frac{n_i(n-n_i)(n+1)}{12}. \quad (8.10)$$

The test statistic

$$H = \sum_{i=1}^k \frac{[R_i - \mu_{R_i}]^2}{\sigma_{R_i}^2} = \sum_{i=1}^k \frac{[R_i - n_i(n+1)/2]^2}{n_i(n-n_i)(n+1)/12} \quad (8.11)$$

will approximately follow a chi-square distribution with $k - 1$ degrees of freedom.

Equation (8.11) can be simplified as

$$H = \frac{12}{n(n+1)} \left(\sum_{i=1}^k \frac{R_i^2}{n_i} \right) - 3(n-1). \quad (8.12)$$

If there are ties, the formula should be further adjusted as

$$H_C = \frac{H}{1 - \sum (t_p^3 - t_p) / (n^3 - n)}, \quad (8.13)$$

where t_p is the number of individuals within p th tied subgroup.

(4) Determine P value and make decision.

Given the value of α , after checking the table of χ^2 distribution with degrees of freedom $k - 1$, if the P -value corresponding to the value of the statistic is less than α , H_0 can be rejected, otherwise, it cannot be rejected.

Example 8.4 14 newborn infants were grouped into four categories according to their mother's smoking habit. (A: smoking more than 20 cigarettes per day; B: smoking less than 20 cigarettes per day; C: ex-smoker; D: never smoke). Their weights are listed in Table 8.9. Now we try to compare the differences of weights in the four groups.

Solution Assume that the shapes of four population distributions are similar.

H_0 : The medians of weights of newborn infants of four categories of mothers with different smoking habits are equal.

Table 8.9 The weights and ranks of newborn infants and their mothers' smoking habit.

Weights x_{ij} (kg)				Ranks r_{ij}			
A	B	C	D	A	B	C	D
2.7	2.9	3.3	3.5	3	4	7	11
2.4	3.2	3.6	3.6	2	5.5	12.5	12.5
2.2	3.2	3.4	3.7	1	5.5	9	14
3.4		3.4		9		9	
n_i				4	3	4	3
R_i				15	15	37.5	37.5

H_1 : The medians of weights of newborn infants of four categories of mothers are not all equal.

- (1) Combine the four samples and rank all observations from smallest to largest. The average rank is shared by each individual in the tied subgroup (see the right of Table 8.9).
- (2) Calculate the rank sums for the four samples, denoted by R_i , $i = 1, 2, 3, 4$ respectively (see the lower part of Table 8.9).
- (3) By (8.12), the test statistic

$$H = \frac{12}{14(14+1)} \left(\frac{15^2}{4} + \frac{15^2}{3} + \frac{37.5^2}{4} + \frac{37.5^2}{3} \right) - 3(14+1) = 9.375.$$

There are three tied subgroups: one with two 3.2, one with three 3.4 and one with two 3.6. The adjusted statistic is

$$H_C = \frac{9.375}{1 - \frac{[(2^3-2)+(3^3-3)+(2^3-2)]}{14^3-14}} = 9.5018.$$

- (4) Determine P value and make decision.

In this case, $k = 4$, $H_C = 9.5018 > \chi_{0.05(3)}^2 = 7.815$, then $P < 0.05$. H_0 can be rejected at the level of $\alpha = 0.05$. We may conclude that the mothers' smoking habit might influence the development of the infants.

8.3.2 Comparison of ordinal data

When the outcome variable is ordinal, the Kruskal–Wallis test for k independent samples is also appropriate for comparing the average levels among groups.

Example 8.5 Alopecia was one of the serious side effects of chemotherapy for cancer. Before a clinical trial, 206 patients with cervical carcinoma were randomly divided into three groups, of which 60 patients in group 1 were planned to take treatment A and 53 patients in group 2 were planned to take treatment B, while 93 patients in group 3 were planned to take treatment C. The degree of hair loss was assessed for each subject after the treatments. The results are listed in Table 8.10. The question is whether the

Table 8.10 Comparison of the degree of hair loss in the three groups.

Degree of hair loss	A	B	C	Total	Range of ranks	Average ranks	Sum of ranks		
							A	B	C
Normal	2	4	13	19	1–19	10	20	40	130
Mild	17	15	22	54	20–73	46.5	790.5	697.5	1023
Moderate	30	28	51	109	74–182	128	3840	3584	6528
Severe	11	6	7	24	183–206	194.5	2139.5	1167	1361.5
Total	60	53	93	206	—	—	6790	5488.5	9042.5

differences of the degree of hair loss in the three groups are statistically significant.

Solution (1) The test hypotheses are

H_0 : The degree of hair loss of the three groups is the same.

H_1 : The degree of hair loss of the three groups is not all the same. $\alpha = 0.05$.

(2) Compute the ranks for the data. First collect all subjects with the same antibody level in the three groups, as shown in Table 8.10. There are 19 with “normal” degree of hair loss who have a rank range of 1–19 and are assigned an average rank of $(1 + 19)/2 = 10$, and similarly for other groups. By doing so, we can get the average rank for the rest of degrees of hair losses, 46.5, 128, 197.5.

(3) Calculate the sums of ranks for the three groups, see in the last few rows of Table 8.10. The sum of ranks for each group equals the sum of the products of the total number and average rank of each degree of hair loss. $R_A = 6790$, $R_B = 5488.5$, $R_C = 9042.5$. According to Eqs. (8.12) and (8.13),

$$H = \frac{12}{206(206 + 1)} \left(\frac{6790^2}{60} + \frac{5488.5^2}{53} + \frac{9042.5^2}{93} \right) - 3(206 + 1) = 2.606,$$

$$H_C = \frac{2.606}{1 - \frac{(19^3 - 19) + (54^3 - 54) + (109^3 - 109) + (24^3 - 24)}{206^3 - 206}} = 3.134.$$

(4) Statistical decision and conclusion. Since $H_c = 3.134$, $\nu = k - 1 = 3 - 1 = 2$, $P = 0.209 > 0.05$, H_0 will not be rejected. We can conclude that the differences of degree of hair loss among the three groups are not statistically significant.

8.3.3 Friedman test for the data from a randomized block design

Just as we need a nonparametric test analog to the parametric one-way analysis of variance, we may find it necessary to have a nonparametric test analog to the parametric two-way analysis of variance. Such a need may arise when the assumptions necessary for parametric analysis of variance are not met, as the measurement scale employed is not very accurate. A test frequently employed under these circumstances is the Friedman two-way analysis of variance for ranks. This test is appropriate whenever the data are arranged in a two-way classification as is given for the randomized block experiment. Let x_{ij} be the observation in i th block of j th treatment group $i = 1, 2, \dots, b$, $j = 1, 2, \dots, k$ and the data format is listed in Table 8.11.

The hypotheses to be tested are

H_0 : The medians of effects in k treatment populations are equal.

H_1 : The medians of effects in k populations are not all equal.

The basic steps are as follows:

- (1) Rank the observations in each block (row) from smallest to largest. If there is a tie, they will be dealt with as before.
- (2) Add up the ranks in each treatment group.

Table 8.11 Data format for the data from a randomized block design.

Blocks	Treatments			
	1	2	...	k
1	x_{11}	x_{12}	...	x_{1k}
2	x_{21}	x_{22}	...	x_{2k}
...
b	x_{b1}	x_{b2}	...	x_{bk}

(3) Calculate the test statistic.

Let r_{ij} be the rank of the j th treatment group in i th block. The rank sum of the i th block is

$$\sum_{j=1}^k r_{ij} = \frac{k(k+1)}{2}.$$

Let R_j be the rank sum of the j th treatment group. The total rank sum of all observations is

$$\sum_{j=1}^k R_j = \frac{bk(k+1)}{2}. \quad (8.14)$$

When H_0 is true, the expectation and variance of R_j are

$$\mu_{R_j} = \frac{b(k+1)}{2}, \quad (8.15)$$

$$\sigma_{R_j}^2 = \frac{b(k^2-1)}{12}. \quad (8.16)$$

It can be proved that when sample size is large enough, the test statistic

$$Z_j = \frac{R_j - \mu_{R_j}}{\sqrt{\sigma_{R_j}^2}} \quad (8.17)$$

approximately follows a standard normal distribution. But with the constraint of formula (8.14), Z_j ($j = 1, 2, \dots, k$) not independent from each other, which means that $\sum_{j=1}^k Z_j^2$ does not follow a χ^2 distribution, then calculate the weighted sum of Z_j^2 ,

$$\chi^2 = \sum_{j=1}^k \left(\frac{k-1}{k} \right) Z_j^2 = \sum_{j=1}^k \frac{[R_j - b(k+1)/2]^2}{kb(k+1)/12}. \quad (8.18)$$

It can be proved, when H_0 is true, that this statistic follows a χ^2 distribution with $k-1$ degrees of freedom.

(4) Statistical decision and conclusion. Given a value of α , if the P -value is less than α , then H_0 can be rejected; otherwise, it cannot be rejected.

Example 8.6 The riboflavin were tested in three samples of cabbage under four test conditions (A , B , C and D). The results are listed in Table 8.12.

Table 8.12 The Riboflavin in cabbages ($\mu\text{g/g}$).

Sample	Test conditions			
	A	B	C	D
1	27.2(2)	24.6(1)	39.5(4)	38.6(3)
2	23.2(1)	24.2(2)	43.1(4)	39.5(3)
3	24.8(2)	22.2(1)	45.2(4)	33.0(3)
R_j	5	4	12	9

Now the question is if the test results are different under different kinds of test conditions.

Solution To test the hypotheses

H_0 : The medians of riboflavin under four test conditions are equal.

H_1 : The medians of riboflavin under four test conditions are not all equal
 $\alpha = 0.05$.

- (1) Rank the observations in each block (row) from smallest to largest. If there is tie, the average rank is shared by each (see the figures in parentheses in Table 8.12).
- (2) Add up the ranks for each test condition (see the last row of Table 8.12).
- (3) Calculate the test statistic, $b = 3, k = 4$,

$$\begin{aligned}
 \chi^2 &= \sum_{j=1}^k \frac{[R_j - b(k+1)/2]^2}{kb(k+1)/12} = \frac{12}{bk(k+1)} \sum_{j=1}^k R_j^2 - 3b(k+1) \\
 &= \frac{12}{3(4)(4+1)} (5^2 + 4^2 + 12^2 + 9^2) - 3(3)(4+1) = 8.2, \\
 &\quad v = k - 1 = 3.
 \end{aligned}$$

- (4) Statistical decision and conclusion. Since $\chi_{0.05}^2 = 7.815 < 8.2$ and $P < 0.05$, H_0 is rejected. We conclude that the results under different conditions are not all equal.

8.3.4 Multiple comparisons of mean ranks in k groups

Like in parametric test for more than two populations, the results of Kruskal-Wallis test or Friedman test only help us to conclude that the locations of

k populations are not all equal but do not help to decide which pairs of populations are different. For latter, multiple comparisons should be done.

For the data of completely randomized design, let \bar{R}_i and \bar{R}_j be the mean ranks of group i and j . The difference between them is $\bar{R}_i - \bar{R}_j$ and its variance is

$$\sigma_{\bar{R}_i - \bar{R}_j}^2 = \frac{n(n+1)}{12} \left[\frac{1}{n_i} + \frac{1}{n_j} \right]. \quad (8.19)$$

In Eq. (8.19), n is the total sample size of k groups, n_i and n_j are the sample sizes of group i and j respectively. The test hypotheses are

H_0 : The medians of group i and j are equal.

H_1 : The medians of group i and j are not equal.

The test statistic is

$$Z_{ij} = \frac{\bar{R}_i - \bar{R}_j}{\sigma_{\bar{R}_i - \bar{R}_j}}. \quad (8.20)$$

Suppose one intends to perform c times of comparison, given the total significance level α , the significance level for each comparison should be adjusted by Bonfferoni method. That is

$$\alpha^* = \frac{\alpha}{c}.$$

If the p -value corresponding to the value of $|Z_{ij}|$ is less than α^* , then H_0 can be rejected; otherwise, it cannot be rejected.

Example 8.7 Perform multiple comparisons for Example 8.4, take the group of non-smoking mother as the reference group.

Solution The mean ranks of observations in the four groups are

$$\begin{aligned} \bar{R}_1 &= 15/4 = 3.75, & \bar{R}_2 &= 15/3 = 5.00, \\ \bar{R}_3 &= 37.5/4 = 9.38, & \bar{R}_4 &= 37.5/3 = 12.50. \end{aligned}$$

In this case, $c = 3$. Given the total $\alpha = 0.05$, then the significance level for each comparison should be

$$\alpha^* = \frac{\alpha}{c} = \frac{0.05}{3} = 0.0167.$$

And the corresponding one-side critical value of the standard normal distribution is $Z_{0.0167} = 2.12$. By (8.20)

$$Z_{1,4} = \frac{\bar{R}_1 - \bar{R}_4}{\sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_1} + \frac{1}{n_4} \right)}} = \frac{3.75 - 12.50}{\sqrt{\frac{14(14+1)}{12} \left(\frac{1}{4} + \frac{1}{3} \right)}} = -2.74,$$

$$Z_{2,4} = \frac{\bar{R}_1 - \bar{R}_4}{\sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_2} + \frac{1}{n_4} \right)}} = \frac{5 - 12.50}{\sqrt{\frac{14(14+1)}{12} \left(\frac{1}{3} + \frac{1}{3} \right)}} = -2.20,$$

$$Z_{3,4} = \frac{\bar{R}_1 - \bar{R}_4}{\sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_4} \right)}} = \frac{9.38 - 12.50}{\sqrt{\frac{14(14+1)}{12} \left(\frac{1}{4} + \frac{1}{3} \right)}} = -0.98.$$

Since $|Z_{1,4}|$ and $|Z_{2,4}|$ are greater than $Z_{0.0187} = 2.12$, we conclude that for smoking mother, the weights of then infants are significantly different from that of non-smoking mothers.

The same procedure may be used for the data from a randomized block design and the only difference is the calculation of variance of the difference between two means of ranks, that is, Eq. (8.19) should be replaced by

$$\sigma_{\bar{R}_i - \bar{R}_j}^2 = \frac{k(k+1)}{6b}. \quad (8.21)$$

8.4 Computerized Experiments

Experiment 8.1 Nonparametric test for the data of paired design The procedure is introduced through Example 8.1.

Lines 01 to 08 in Program 8.1, input the data to SAS database NPAR1 and calculate the difference (D) and absolute difference (BASD). Lines 10 to 19 rank the differences, attach their original signs and print the ranking results. Lines 22 and 23 calculate the test statistic by procedure UNIVARIATE in SAS. If the t -test is used for this example, the results are $t = -0.9281$, $P = 0.3733$. By Wilcoxon's signed-rank test, the test statistic is -8.5 , $P = 0.4756$.

Experiment 8.2 Comparison between parametric test and nonparametric test through the data from a normal distribution Two samples are

Program 8.1 Wilcoxon's signed-rank test.

Line	Program	Line	Program
01	DATA NPAR1;	12	RANKS ABSR;
02	INPUT X Y @@;	13	DATA B;
03	D=Y-X;	14	SET RA;
04	IF D=0 THEN ABSD=.;	15	IF D<0 THEN DRANK=-ABSR;
05	ELSE ABSD=ABS (D) ;	16	IF D=0 THEN DRANK=. ;
06	CARDS;	17	IF D>0 THEN DRANK=ABSR;
07	86 88 71 77 77 76 68 64 91 96 72 72	18	PROC PRINT DATA=B;
08	77 65 91 90 70 65 71 80 88 81 87 72	19	VAR X Y D ABSD ABSR
09	;		DRANK;
10	PROC RANK DATA=NPARI	20	PROC UNIVARIATE;
	TIES=MEAN OUT=RA;	21	VAR D;
11	VAR ABSD;	22	RUN;

Program 8.2 T-test and rank-sum test for comparison of two samples.

Line	Program	Line	Program
01	DATA NPAR2;	11	VAR X;
02	INPUT GRP \$ X @@;	12	PROC RANK DATA=NPARI OUT=A
03	CARDS;		TIES=MEAN;
04	A 25 A 34 A 44 A 46 A 46 B 15 B 15 B 16 B 17	13	VAR X;
05	B 19 B 21 B 21 B 23 B 25 B 27 B 28 B 28 B 30	14	RANKS RANKX;
06	B 35;	15	PROC PRINT DATA=A;
07	PROC UNIVARIATE	16	PROC NPARIWAY DATA=NPARI
	PLOT NORMAL;		ANOVA WILCOXON MEDIAN;
08	VAR X;	17	VAR X;
09	PROC TTEST COCHRAN;	18	CLASS GRP;
10	CLASS GRP;	19	RUN;

drawn from the same normal distribution and then parametric and nonparametric procedures are used to test the null hypothesis: the population means are equal.

In Program 8.2, Lines 01 to 06 input the data of Example 8.2 into SAS database NPAR2. Lines 07 and 08 test the normality condition and the results of test are $w_A = 0.8257$, $P_A = 0.1292$; $w_B = 0.9461$, $P_B = 0.5024$. Lines 09 to 11, perform a t -test for two independent samples and the results

Program 8.3 Generate skew distributed data and compare parametric and non-parametric tests.

Line	Program	Line	Program
01	DATA A;	11	VAR X;
02	DO I=1 TO 40;	12	PROC TTEST DATA=A
03	X=EXP (RANNOR (286455)) ;		COCHRAN;
04	IF I<21 THEN GRP='A';	13	CLASS GRP;
05	ELSE GRP='B';	14	VAR X;
06	OUTPUT;	15	PROC NPARIWAY DATA=A
07	END;	16	ANOVA WILCOXON MEDIAN;
08	PROC PRINT;	17	VAR X;
09	PROC UNIVARIATE DATA=A	18	CLASS GRP;
10	PLOT NORMAL;	19	RUN;

are $t = 4.3982$, $P = 0.0004$. Lines 12 to 15 calculate the rank sum and print the results. Lines 16 to 19 perform a nonparametric test and the results are $Z = 2.5980$, $P = 0.0094$.

Experiment 8.3 Comparison between parametric test and nonparametric test through the data from a skewed distribution population In Program 8.3, lines 01 to 07 produce 40 values from an exponential distribution, and divide them into groups *A* and *B*. Lines 09 to 11, test the normality condition. Lines 12 to 14, perform a *t*-test and lines 15 to 19, perform a Wilcoxon's rank-sum test.

Experiment 8.4 Nonparametric procedures for comparison among multiple samples In Program 8.4, lines 01 to 09 input the data of Example 8.3 into SAS database NPAR1. Lines 10 to 12 perform an analysis of variance by SAS procedure NPARIWAY and lines 13 to 15 calculate the Kruskal–Wallis test statistic, $H = 9.50$, $P = 0.0233$. Lines 16 to 18 rank the observations and then perform multiple comparisons.

8.5 Practice and Experiments

1. When will the nonparametric methods be adopted? When will the parametric methods be adopted? Why? What are the advantages and the disadvantages of the nonparametric methods?

Program 8.4 Rank-sum test for the data from completely randomized designed experiment.

Line	Program	Line	Program
01	DATA NPAR1;	15	CLASS GRP;
02	INPUT NO GRP\$ X@@;	16	PROC RANK DATA=NPARI
03	CARDS;	17	OUT=A;
04	1 1 2.7 2 1 2.4 3 1 2.2	18	VAR X;
05	4 1 3.4 5 2 2.9 6 2 3.2	19	RANKS R;
06	7 2 3.2 8 3 3.3 9 3 3.6	20	PROC TABULATE;
07	10 3 3.4 11 3 3.4 12 4 3.5	21	CLASS GRP NO;
08	13 4 3.6 14 4 3.7	22	VAR X R;
09	;	23	TABLE GRP*NO, X R;
10	PROC NPARIWAY ANOVA;	24	PROC ANOVA ;
11	CLASS GRP;	25	CLASS GRP;
12	VAR X;	26	MODEL R=GRP;
13	PROC NPARIWAY WILCOXON;	27	MEANS GRP/BON;
14	VAR X;	28	RUN;

Table 8.13 The serum total cholesterol pre- and post-treatment (mmol/L).

No. of cases	Before	After	No. of cases	Before	After
1	560	223	2	975	220
3	550	205	4	720	235
5	742	236	6	470	220
7	450	230	8	460	240
9	422	190	10	276	198
11	280	220	12	170	198
13	250	190	14	426	188
15	210	240	16	260	186
17	230	254			

- 2. 17 cases of chronic nephritis patients receive in a period of time, hormone and immunosuppressant. Before and after treatment, their serum total cholesterol (mmol/L) is measured. The results are listed in Table 8.13. Try to evaluate the effect of the treatment on serum total cholesterol.
- 3. The serum albumin (g/L) of 14 normal adults and 13 in-patients are measured. The results are listed in Table 8.14. Do you think that the difference of medians of serum albumin between the two groups is statistically significant? And why?

Table 8.14 The results of serum albumin test of 14 normal adults and 13 in-patients (g/L).

Normal	24	35	31	40	42	34	45	30	32	35	38	39	40	35
In-patient	15	20	34	17	20	38	35	15	18	20	15	13	31	

Table 8.15 The analgesic effect of Neiguan and Zusanli.

Acupuncture point	Effect				
	Very good	Good	Median	Bad	Very bad
Neiguan	25	20	16	12	10
Zusanli	28	30	44	33	30

Table 8.16 The survival days of rats without hypophysis with different doses of ACH.

Dose A (no ACH)	Dose B	Dose C	Dose D
3	2	4	12
2	1	3	13
1	3	4	7
2	5	4	6
3	7	6	8
5	14	5	19
4	8	4	20
2	15	3	5
2	3	5	2
3	4	5	12

- In order to compare the analgesic effects of two acupuncture points (Neiguan and Zusanli), a clinical trial was conducted in a hospital of traditional Chinese medicine. The results were listed in Table 8.15. Now the question is which acupuncture point has a better analgesic effect.
- The rats are taken off their hypophysis and then are randomly assigned into four groups to receive different doses of adrenocortical hormone (ACH). The survival days of rats in different groups are listed in Table 8.16. Try to test the hypothesis that the survival time of the rats without hypophysis is independent of the level of ACH.

Table 8.17 The days with different air qualities in three cities.

Cities	Air qualities				
	Good	Moderate	Unhealthy for sensitive group	Unhealthy	Very unhealthy
A	9	15	42	60	74
B	32	29	55	36	48
C	69	46	33	21	31

Table 8.18 Reaction rates (%) of 8 subjects to different audio frequency stimulations.

No. of subjects	Frequency A	Frequency B	Frequency C	Frequency D
1	8.4	9.6	9.8	11.7
2	11.6	12.7	11.8	12.0
3	9.4	9.1	10.4	9.8
4	9.8	8.7	9.9	12.0
5	8.3	8.0	8.6	8.6
6	8.6	9.8	9.6	10.6
7	8.9	9.0	10.6	11.4
8	7.8	8.2	8.5	10.8

- 6. According to the data in Table 8.17, try to apply certain hypothesis tests to compare the air qualities of three cities.
- 7. The experimental subjects are exposed to four different audio frequency stimulations respectively under the same laboratory condition. The reaction rates of them are listed in Table 8.18. Now try to analyze the difference of reaction rates to different frequency stimulations.

(1st edn. Qing Liu, Jiqian Fang; 2nd edn. Jinxin Zhang, Jiqian Fang)

Chapter 9

Simple Linear Correlation

In the previous chapters, some statistical methods dealing with a continuous variable have been introduced. Very often we need to study the relationship between two random variables, such as blood pressure and body mass index. If the two variables vary together without differentiation of dominant and subordinate, the analysis of simple linear correlation may be useful, which will be introduced in this chapter. If we assume one variable as the response variable (dependent variable) and another as an explanatory variable (independent variable), the regression method is more suitable, which will be introduced in the next chapter.

9.1 Concept of Correlation

9.1.1 *Independent random sample of a bivariate normal distribution*

Let $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ be an independent sample from a joint distribution of a pair of variables (X, Y) . In practice, X and Y may be the measures of two characteristics of the same individual (like the age of a person and his systolic blood pressure), or may be the same variable of paired individuals (like the IQ scores of twins). The question here is the association between X and Y . If larger values of Y tend to correspond to larger values of X , we say that X and Y are positively correlated; otherwise, if larger values of Y tend to correspond to smaller values of X , we say that X and Y are negatively correlated.

Example 9.1 To explore the linear correlation between systolic blood pressure and diastolic blood pressure, the systolic and diastolic blood pressures of 665 girls aged from 6 to 10 years were recorded (mmHg).

Table 9.1 Heights (cm) of 20 pairs of father and son.

No.	1	2	3	4	5	6	7	8	9	10
Father's height, X	150	153	155	158	161	164	165	167	168	169
Son's height, Y	159	157	163	166	169	170	169	167	169	170
No.	11	12	13	14	15	16	17	18	19	20
Father's height, X	170	171	172	174	175	177	178	181	183	185
Son's height, Y	173	170	170	176	178	174	173	178	176	180

Here the diastolic blood pressure and systolic blood pressure of the same person can be regarded as two random variables X and Y , and if the 665 girls were randomly selected from a population, then their records can be used to explore the association between the two measures.

Example 9.2 To explore the linear correlation between the heights of father and son, 20 graduate male students were randomly selected from a name list in a high school. The heights (cm) of fathers and sons were measured, and given in Table 9.1.

The above description of Example 9.2 has mentioned that the graduate male students were “randomly selected”. However, if there were close relatives, brothers, and even twins being included in the sample, then it will not be an independent sample for analysis of linear correlation. Here we assume that before sampling, those possible relatives and brothers have been removed from the name list. Then the individuals in the sample can be regarded as independent.

9.1.2 Scatter diagram

The most simple and intuitive way to explore correlation between two random variables is to plot a scatter diagram, where two variables are expressed by two coordinate axes; n pairs of observed data were expressed by n points in the coordinate plane.

The $n = 665$ records in Example 9.1 can be plotted in the $X - Y$ plane as in Fig. 9.1, where X refers to diastolic pressure, Y refers to systolic pressure, and each point refers to one girl. We can see that larger X values tend to associate with larger Y values and vice versa. We say that diastolic pressure and systolic pressure are positively correlated.

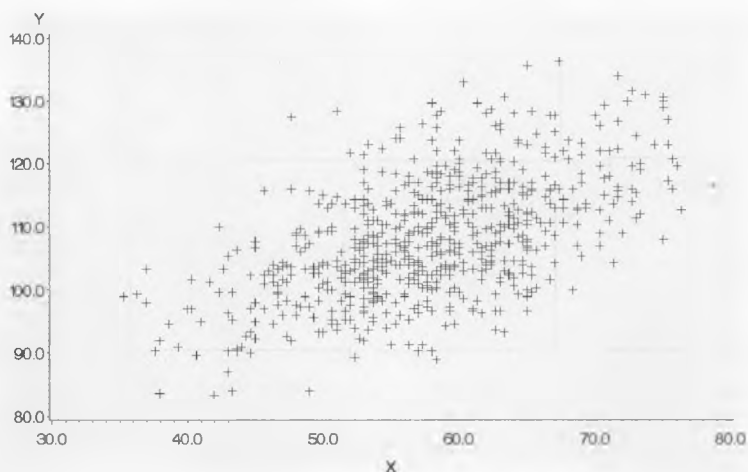


Fig. 9.1 Scatter diagram of systolic and diastolic blood pressures (mmHg) of 665 girls aged within 6–10 years old.

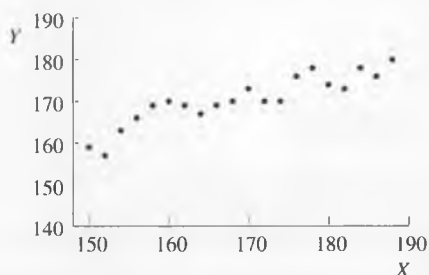


Fig. 9.2 Scatter diagram for Example 9.2.

The scatter diagram for the 20 pairs of the height data is given in Fig. 9.2. One can see that it is not necessary that a taller father with have a taller son, but as a whole taller fathers tend to have taller sons.

Several typical patterns of scatter diagrams are showed in Fig. 9.3(a) and (c) show a linearly increasing tendency of Y with the increasing of X , which is subject to positive correlation; While (b) and (d) show a linearly decreasing tendency, which is subject to negative correlation; (e), (f) and (g) show that there is no any association between X and Y ; (h) shows a tendency of curvilinear association between X and Y . (e), (f) and (g) as

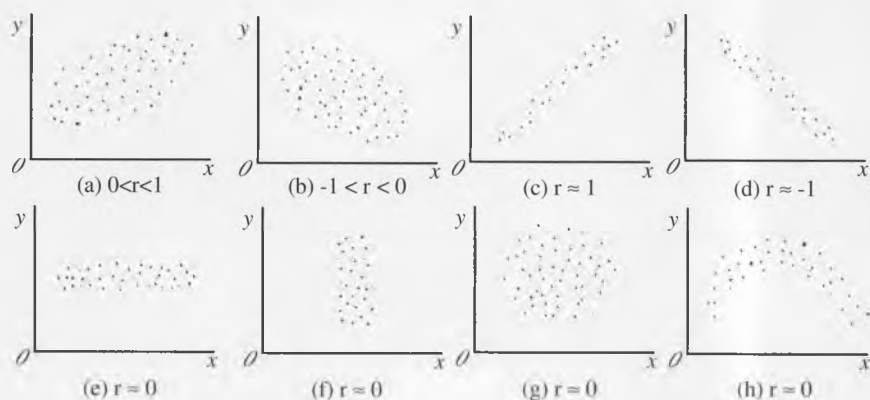


Fig. 9.3 The typical scatter diagrams.

well as (h) are all subject to “not linearly correlated”. Therefore, whenever people speak of “none of linear correlation”, it is necessary to clarify whether there is “no any association” or there is actually “curvilinear association”.

9.2 Correlation Coefficient

9.2.1 Population correlation coefficient

It is desirable to have a measure for the association, with its sign indicating whether the association is negative or positive and its absolute value between 0 and 1 indicating the strength of association. There are other properties that we require on such a measure. We like it to be invariant of (i.e. not affected by) the general level of the variables of X and Y , nor the measurement scale of X and Y . Naturally, we start searching for a coefficient based on the standardized variables $(X - \mu_x)/\sigma_x$ and $(Y - \mu_y)/\sigma_y$, where μ_x and μ_y are the means of X and Y , σ_x and σ_y are the standard derivations of X and Y . The mean of product of the two standardized variables is called the Pearson's product-moment linear correlation coefficient, or population correlation coefficient, denoted by

$$\rho = E \left[\left(\frac{X - \mu_x}{\sigma_x} \right) \left(\frac{Y - \mu_y}{\sigma_y} \right) \right]. \quad (9.1)$$

(9.1) can also be written as

$$\rho = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \quad (9.2)$$

where the numerator

$$\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)] \quad (9.3)$$

is called population covariance between X and Y .

If $\rho = 0$, it is called no linear correlation or null correlation between X and Y ; $\rho > 0$, positive correlation; $\rho < 0$, negative correlation; $\rho = 1$ or $\rho = -1$, complete correlation, which is extremely rare in real life.

9.2.2 Sample correlation coefficient

Given an independent sample $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ with sample means \bar{x} and \bar{y} , replacing the numerator and denominator on the right-hand side of (9.2) with their sample estimators, one can get the sample correlation coefficient, denoted with r ,

$$r = \frac{S_{xy}}{S_x S_y}, \quad (9.4)$$

where S_{xy} is the sample covariance between X and Y ,

$$S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}. \quad (9.5)$$

To an individual, a positive product $(x_i - \bar{x})(y_i - \bar{y})$ implies that both X and Y of this individual are located in the same direction against the means \bar{x} and \bar{y} ; a negative product $(x_i - \bar{x})(y_i - \bar{y})$ implies that X and Y of this individual are located in different directions against the means \bar{x} and \bar{y} .

A positive sum of the product $(x_i - \bar{x})(y_i - \bar{y})$ implies that most of the individuals having their X and Y located in the same direction against the means \bar{x} and \bar{y} such that we say X and Y are positively correlated. Similarly, a negative sum of the product $(x_i - \bar{x})(y_i - \bar{y})$ implies that most of the individuals having their X and Y located in different direction against the means \bar{x} and \bar{y} such that we say X and Y are negatively correlated. Contrarily, a zero sum of the product $(x_i - \bar{x})(y_i - \bar{y})$ implies that about half of individuals having their X and Y located in the same direction against the means \bar{x} and \bar{y} , and another half of individuals having their X and Y located

in different directions against the means \bar{x} and \bar{y} , that is, as a whole, the locations of X and Y against the means \bar{x} and \bar{y} vary randomly such that we say X and Y are not linearly correlated.

9.2.3 Calculation of sample correlation coefficient

Denote

$$l_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad l_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2,$$

$$l_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Since

$$S_x^2 = \frac{l_{xx}}{n-1}, \quad S_y^2 = \frac{l_{yy}}{n-1}, \quad S_{xy} = \frac{l_{xy}}{n-1},$$

the sample correlation coefficient can be calculated by

$$r = \frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}. \quad (9.6)$$

For convenience, l_{xx} , l_{yy} and l_{xy} can be calculated by the following formulas

$$l_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2,$$

$$l_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2,$$

$$l_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)$$

$$= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}. \quad (9.7)$$

Example 9.3 (Cont'd of Example 9.2) Calculate the sample correlation coefficient between the heights of father and son.

Solution $\sum_{i=1}^n x_i = 3376$, $\sum_{i=1}^n y_i = 3407$, $n = 20$, $\sum_{i=1}^n x_i^2 = 571728$, $\sum_{i=1}^n y_i^2 = 581081$, $\sum_{i=1}^n x_i y_i = 576161$. By (9.7),

$$l_{xx} = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 = 571728 - \frac{3376^2}{20} = 1859.2,$$

$$l_{yy} = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 = 581081 - \frac{3407^2}{20} = 698.55,$$

$$\begin{aligned} l_{xy} &= \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) \\ &= 576161 - \frac{(3376)(3407)}{20} = 1059.4, \end{aligned}$$

$$r = \frac{l_{xy}}{\sqrt{l_{xx} l_{yy}}} = \frac{1059.4}{\sqrt{1859.2 \times 698.55}} = 0.9296.$$

9.3 Inference on Correlation Coefficient

In practice, the population variances σ_x^2 , σ_y^2 and population covariance σ_{xy} are unknown. Thus the sample correlation coefficient r is only an estimate of the population correlation coefficient ρ . Usually the sample correlation coefficients based on different samples are different from each other. This shows that the sample correlation coefficient is a random variable in nature, denoted by R , while the calculated value r is just one observation of R . In order to clarify whether the population correlation coefficient is really not zero and to estimate the actual level of the population correlation coefficient with a confidence interval, it is necessary to work out statistical inference right after a sample correlation coefficient r is obtained.

It has to be noted that the following inference is based on the assumption that X and Y follow a bivariate normal distribution. While for the definition and calculation of correlation coefficient, this is not necessary.

9.3.1 Hypothesis test

First we need to know whether the population correlation coefficient ρ is equal to zero or not. If $\rho \neq 0$, we say X and Y are linearly correlated.

However, $r \neq 0$ does not necessarily mean $\rho \neq 0$, because even a sample comes from a population with $\rho = 0$, it is still possible to have a sample correlation coefficient $r \neq 0$. Therefore, we need to test

$$H_0 : \rho = 0, \quad H_1 : \rho \neq 0.$$

When sample size $n \leq 52$, one can use Table 8 in appendix II to find a p -value corresponding to the sample correlation coefficient r directly.

Usually, a t test is also available for this inferential analysis. When H_0 is true, the statistic

$$t_R = \frac{R}{S_R} \sim t \text{ dist. } \nu = n - 2, \quad (9.8a)$$

where

$$S_R \approx \sqrt{\frac{1 - r^2}{n - 2}} \quad (9.9a)$$

is the standard deviation of sample correlation coefficient R , it is also called standard error of R .

Example 9.4 (Cont'd of Example 9.3) After getting $r = 0.9296$, it is required to test whether r is statistically significant.

Solution To test $H_0 : \rho = 0$, $H_1 : \rho \neq 0$, by (9.8a) and (9.9a), we have

$$S_R \approx \sqrt{\frac{1 - 0.9296^2}{20 - 2}},$$

$$t_r = \frac{r}{S_R} = \frac{0.9296}{\sqrt{\frac{1 - 0.9296^2}{20 - 2}}} = 10.7.$$

From the table for t distribution, we have $t_{0.001, 18} = 3.922$. Obviously, $|t_r| > 3.922$, $P < 0.001$ so that H_0 is rejected at the level of $\alpha = 0.001$. It could be concluded that there is positive correlation between the heights of father and son.

In fact, if Table 8 in Appendix II is checked directly, we have $r_{0.001, 18} = 0.679$, $|r| > r_{0.001, 18}$, $P < 0.001$, resulting in the same conclusion.

9.3.2 Interval estimation

The hypothesis test for $H_0 : \rho = 0$ is only to answer the question whether the linear correlation exists. Once H_0 is rejected, one would further like to know the strength of the correlation, that is, a confidence interval of the correlation coefficient ρ is required.

It is known that when H_0 is true, the statistic

$$Z = \tanh^{-1} r \quad (9.10)$$

approximately follows a normal distribution

$$N\left(\tanh^{-1} \rho, \frac{1}{\sqrt{(n-3)}}\right).$$

Therefore, the $(1 - \alpha)$ confidence interval of $\tanh^{-1} \rho$ is approximately

$$\tanh^{-1} \rho : \left(\tanh^{-1} r - Z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n-3}}, \tanh^{-1} r + Z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n-3}} \right) \quad (9.11a)$$

or

$$\tanh^{-1} \rho : \tanh^{-1} r \pm Z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n-3}}. \quad (9.11b)$$

Taking a transformation of $\tanh(*)$ for (9.11), one can get a $(1 - \alpha)$ confidence interval for ρ without difficult.

Remark. The function $\tanh(*)$ is called hyperbolic tangent, and $\tanh^{-1}(*)$ is its inverse which is defined as

$$\tanh^{-1} R = \frac{1}{2} \ln \left(\frac{1+R}{1-R} \right) \quad (9.12)$$

and

$$\tanh Z = \frac{\exp(2Z) - 1}{\exp(2Z) + 1}. \quad (9.13)$$

Example 9.5 (Cont'd of Example 9.4) After getting $r = 0.9296$, find a 95% confidence interval for the population correlation coefficient ρ .

Solution By (9.12),

$$\tanh^{-1} r = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) = \frac{1}{2} \ln \left(\frac{1+0.9296}{1-0.9296} \right) = 1.6554.$$

By (9.11b), the 95% confidence interval of $\tanh^{-1} \rho$ is

$$\tanh^{-1} r \pm Z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n-3}} = 1.6554 \pm 1.96 \frac{1}{\sqrt{20-3}} = (1.1800, 2.1308).$$

Take a transformation of $\tanh(*)$ for it,

$$\tanh 1.1800 = \frac{\exp(2 \times 1.1800) - 1}{\exp(2 \times 1.1800) + 1} = 0.8275,$$

$$\tanh 2.1308 = \frac{\exp(2 \times 2.1308) - 1}{\exp(2 \times 2.1308) + 1} = 0.9722.$$

Finally, we conclude that the 95% confidence interval of correlation coefficient between the heights of father and son is (0.8275, 0.9722).

9.4 Rank Correlation

The above-mentioned inference on Pearson's product-moment linear correlation coefficient requires a pre-requisite of bivariate normal distribution. However, in practical researches the raw data might not follow a normal distribution, or even the distribution is unknown, or sometimes the data are not precisely measured (for instance, limited by the sensitivity of the instrument, the concentration of certain ion is reported as " $<0.001 \mu\text{g/ml}$ "), or X and/or Y themselves are ordinal variables. In those cases, the rank correlation can be used to describe the association between two random variables, including strength and direction. It is a kind of non-parametric statistical methods based on rank. Here we would only introduce the commonly applied Spearman's rank correlation coefficient.

9.4.1 Spearman's rank correlation coefficient

Assume X and Y are continuous variables or ordinal variables and a random sample with n pairs of observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ are available.

Rearrange the queues of x_1, x_2, \dots, x_n and y_1, y_2, \dots, y_n respectively according to their values from minimum to maximum getting the rank p_i for x_i and q_i for $y_i, i = 1, 2, \dots, n$. (In case there is a tie, the average rank will be shared. See Example 10.1). Denote $d_i = p_i - q_i$. The Spearman's rank correlation coefficient is defined as

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}. \quad (9.14)$$

Similar to the Pearson's product-moment correlation coefficient, r_s can reflect the association between two random variables; it is a sample estimation of the population rank correlation coefficient ρ_s ; both ρ_s and r_s are free of unit, and take values between -1 and 1 .

From (9.14), one can see that the basic idea of rank correlation is focused on the extent of consistency between p_i and q_i , which is indicated by $d_i = p_i - q_i$. Since d_i could be positive as well as negative, we rather like to use $\sum d_i^2$ to reflect the extent of un-consistence between p_i and q_i rather than $\sum d_i$. If $\sum d_i^2 = 0$, they are positively correlated with an extreme strength; if $\sum d_i^2$ reaches maximum, they are negatively correlated in either an extreme strength; the value of $\sum d_i^2$ is related to the strength of correlation as showed in (9.14).

Similarly, after calculating the rank correlation coefficient, we also need to work on a hypothesis test

$$H_0 : \rho_s = 0, \quad H_1 : \rho_s \neq 0.$$

When the sample size $n \leq 50$, Table 9 in Appendix II can be used directly to get the critical value for r_s . If $n > 50$, (9.8a) and (9.9a) can still be used as long as the R there being replaced by R_s . That is, when $H_0 : \rho_s = 0$ is true,

$$t_R = \frac{R_s}{S_R} \sim t \text{ dist. } \nu = n - 2, \quad (9.8b)$$

where

$$S_R \approx \sqrt{\frac{1 - r_s^2}{n - 2}}. \quad (9.9b)$$

Example 9.6 In an etiology study on liver cancer, data on liver-cancer-specific death rate ($Y, 1/10^5$) and the relative content of aflatoxin (X) in

Table 9.2 Calculation of rank correlation coefficient for Example 9.6.

No. (1)	Aflatoxin		Liver-cancer-specific death rate		$d = p - q$ (6) = (3) - (5)	d^2 (7) = (6) ²
	X (2)	Rank p (3)	Y (1/10 ⁵) (4)	Rank q (5)		
1	0.7	1	21.5	3	-2	4
2	1.0	2	18.9	2	0	0
3	1.7	3	14.4	1	2	4
4	3.7	4	46.5	7	-3	9
5	4.0	5	27.3	4	1	1
6	5.1	6	64.6	9	-3	9
7	5.5	7	46.3	6	1	1
8	5.7	8	34.2	5	3	9
9	5.9	9	77.6	10	-1	1
10	10.0	10	55.1	8	2	4
Total						42

certain food for ten countries have been collected, which are given in the columns (2) and (4) in Table 9.2. Calculate the rank correlation between the two variables.

Solution The test $H_0 : \rho_s = 0$, $H_1 : \rho_s \neq 0$ can be worked out from Table 9.2. The raw data for X and Y are put in columns (2) and (4), their ranks are listed in columns (3) and (5). Columns (6) and (7) are the calculations for $d_i = p_i - q_i$ and d_i^2 . Put the total of column (7) into (9.14), we have

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} = 1 - \frac{6(42)}{10(10^2 - 1)} = 0.7455.$$

From Table 14 in Appendix II, we have $0.01 < P < 0.02$ so that H_0 is rejected at the level of $\alpha = 0.05$. It can be concluded that there is positive correlation between the liver-cancer-specific death rate and the content of aflatoxin in certain food.

9.4.2 When there are more ties in rank

There is an alternative way to calculate r_s . Putting the ranks p_i and q_i into formulas (9.6) and (9.7) to replace the raw data x_i and y_i , as a result, the product-moment correlation coefficient based on p_i and q_i can be regarded

as the rank correlation coefficient r_s . In fact, when there is no tie among x_1, x_2, \dots, x_n and among y_1, y_2, \dots, y_n , the result through such a way is exactly equal to that through (9.14). When there are more ties among x_1, x_2, \dots, x_n and y_1, y_2, \dots, y_n , formula (9.14) will no longer work well and this alternative method is recommended.

9.5 Caution in Analysis of Linear Correlation

9.5.1 The importance of scatter diagram

Before any analysis of linear correlation, a scatter diagram is always plotted, which might provide valuable guide for further appropriate analysis.

First of all, the association between two variables is not always subject to a linear association. For instance, people with higher blood pressure and with lower blood pressure tend to have higher death rates, while those with medium blood pressure tend to have lower death rates, hence the scatter diagram between death rate and blood pressure looks like (h) of Fig. 9.3 and it reminds us that the analysis of linear correlation is inappropriate for such case.

Secondly, if the scatter diagram shows that the distribution of X and Y does not look like a bivariate normal distribution, the rank correlation should be considered rather than the product moment correlation.

Furthermore, one should be careful when a few outliers appear in the scatter diagram like that in (a) of Fig. 9.4. In case it is possible, some replication in sampling around the outliers is helpful in exploring the data structure; at least, the preliminary records should be doubly checked. It is not allowed to revise or remove any figures unless there is strong evidence

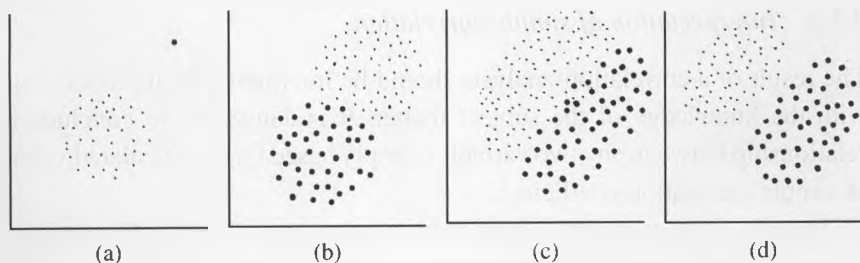


Fig. 9.4 Caution in analysis of correlation. (a) Outlier; (b), (c) and (d) Stratified situation.

to show that the outliers are surely due to a mistake. The rank correlation can help to reduce the impact of “outliers” without any change of the records.

9.5.2 *No correlation for non-random sample*

In some occasions, one variable randomly varies but the values of another are set by the researcher. For instance, to explore the dose response relationship, k dosages are given by the researcher according to the research protocol; and again, to explore the appropriate temperature for the yield of a chemistry reaction, the k temperatures are selected by the researcher according to the research needs. In such cases, although people can still calculate by the formula of correlation coefficient, the result does not well represent the population correlation coefficient because the data are not a random sample. In fact, the result will vary with the range of values selected by the researcher.

9.5.3 *Caution about pooling data*

As showed in (b) of Fig. 9.4, the two variables do not correlate in each of the two strata, but after pooling, people are misled to a correlation by faint; in (c) of Fig. 9.4, the two variables do correlate in each of the two strata, but after pooling, people are misled to a null correlation by faint; in (d) of Fig. 9.4, the two variables positively correlate in each of the two strata, but after pooling, people are misled to a negative correlation by faint. In general, pooling is rational only when the relationships in each of the strata would not be distorted after pooling, otherwise, the stratified analyses are recommended.

9.5.4 *Interpretation of simple correlation*

The result of a correlation analysis should be interpreted by incorporating with the knowledge of the subject matter. It is dangerous to conclude a relationship between any two variables simply based on the calculated value of simple correlation coefficient.

Here is a story recently happened. A man got a son at the autumn of his life. He planted a small tree to accompany his son and measured both the heights of his son and the tree every month, denoted with x_i and

$y_i, i = 1, 2, \dots$ respectively. After n months, he calculated a correlation coefficient and a hypothesis test showed that the correlation coefficient is statistically significant so he told us that there is a strong relationship between his son and the tree. Is it true? In fact, the two heights have no relationship; the correlation coefficient was caused by their links with the factor of "time".

Even though sometimes the correlation between two variables is rational in terms of the subject matter, rather than causation, it is just a kind of numerical association and the two variables are not at dominant versus subordinate positions. Because of this, the correlation discussed in this chapter is called simple correlation.

9.6 Computerized Experiments

Experiment 9.1 Sampling experiment for null correlation Randomly generate 100 independent "observations" from a normal distribution $N(0, 1)$; every two successive ones form a pair such that we can have 50 pairs of the "data" for X and Y . Plot a scatter diagram, calculate a correlation coefficient and test whether the population correlation coefficient is zero. Repeat the above process (not including the scatter diagram) for 100 times. Sum up the total number of rejecting the null hypothesis, and discuss the indication of this summarized number.

Program 9.1 Sampling experiment for null correlation.

Line	Program	Line	Program
01	DATA A;	07	PROC GPLOT;
02	DO I=1 TO 50 BY 1;	08	PLOT Y*X;
03	X=RANNOR(0);	09	PROC CORR;
04	Y=RANNOR(0);	10	VAR X Y;
05	OUTPUT;	11	RUN;
06	END;		

Lines 02–06 form a cycle to generate 50 pairs of random number, where lines 3 and 4 generate normal variables X and Y respectively, and line 5 outputs the "data". Lines 07 and 08 plot a scatter diagram. Lines 09 and 10 calculate the correlation coefficient and test the hypothesis of null correlation.

Experiment 9.2 Sampling experiment for nonzero correlation Randomly generate the values of X from the normal distribution $N(0, 1)$, and then generate the corresponding values of Y from the normal distribution $N(X, 1)$ so that the distribution of Y depends on the value of X . Repeatedly generate 30 pairs of values for X and Y ; plot a scatter diagram, calculate the correlation coefficient and test whether the population correlation coefficient is zero. Repeat the above process (not including the scatter diagram) for 100 times. Sum up the total number of rejecting the null hypothesis, and discuss the indication of this summarized number.

Program 9.1 can still be used, only line 02 is changed as “DO I=1 TO 30 BY 1”, and line 04 is changed as “Y=X + RANNOR(0)”.

9.7 Practice and Experiments

1. To explore the relationship between the level of blood sugar (mmol/L) and that of insulin (mU/L), 20 patients with diabetes were recruited from a hospital in Taiyuan, getting a correlation coefficient $r = -0.523$ (two-side, $0.01 < P < 0.02$); and meanwhile 60 patients with the same type of diabetes were recruited from several hospitals in Guangzhou, getting a correlation coefficient $r = -0.456$ (two-side, $P < 0.001$). Based on the size of P -values, the researcher concluded that the population correlation among the patients in Guangzhou was rather stronger than that in Taiyuan. Give comments on this.
2. If someone would like to work on pooled data for the above study, taking the plots in Fig. 9.4 as reference, discuss what should be considered.
3. If the liver-cancer-specific death rates (Y) in Table 9.2 were ranked from the maximum to the minimum, what will happen to the rank correlation coefficient? If one directly calculates the product moment correlation based on the ranks, what will be the possible result?
4. After calculating the product moment correlation based on the ranks, is it allowed to use the Table 8 of Appendix II for product moment correlation coefficient to find the critical value?
5. Randomly selected 18 students from a high school, the intelligent quotients (IQ) were measured at the end of the year. The results are given in

Table 9.3 The IQs and scores of mathematics and literatures of 18 high school students.

No.	1	2	3	4	5	6	7	8	9
Score of mathematics, X	78	84	61	52	93	89	98	98	65
Score of literature, Y	83	76	70	58	82	78	89	95	61
IQ, Z	95	100	100	75	105	97	110	120	76
No.	10	11	12	13	14	15	16	17	18
Score of mathematics, X	73	48	45	67	75	95	88	99	81
Score of literature, Y	75	53	43	70	78	97	92	92	88
IQ, Z	92	61	60	88	96	125	113	126	102

Table 9.3 incorporating their scores of mathematics and literature of that year.

- (1) Calculate the correlation coefficients between the score of mathematics and IQ, between the score of literature and IQ, and between the scores of mathematics and literature;
- (2) Work out hypothesis tests to see if the population correlation coefficients are significantly different from 0;
- (3) Can we say that good at mathematics might be caused by good at literature, vice versa?

(1st edn. Jiqian Fang, Kai Ng; 2nd edn. Jinxin Zhang, Jiqian Fang)



Chapter 10

Simple Linear Regression

The concept of linear correlation introduced in Chap. 9 is used to describe the extent of linear association between two random variables X and Y , of which both play an equal role, indifference in dominant and subordinate. Furthermore, the researchers may be rather interested in how the value of one variable is affected by that of the other such as how the death rate of animal depends on the drug dosage. The statistical method to explore linear dependence quantitatively between two continuous variables is called simple linear regression, or simple regression for short. Here the two variables play different roles. One is called independent variable or explanatory variable, usually denoted by X , of which the values could be set by the researcher or could be a random variable; another is called dependent variable or response variable, usually denoted by Y , of which the values could randomly follow certain rule once the value of X is given. If the “rule” is described as an equation, one can predict the value of Y corresponding to the given value of X such as predicting the infant weight by age in months or predict the body surface area by height.

10.1 Statistical Description of Linear Regression

10.1.1 *Linear regression equation*

English geneticist Francis Galton (1889) and his students K. Pearson and A. Lee (1903) noticed an interesting phenomenon, so-called “regression to the mean”, that the sons of taller fathers tend to be tall, and the sons of shorter fathers tend to be short, but their heights tend to be closer to the average level of their fathers. It is indeed imaginable. Otherwise, the height will be further away from the average level generation by generation, resulting in a

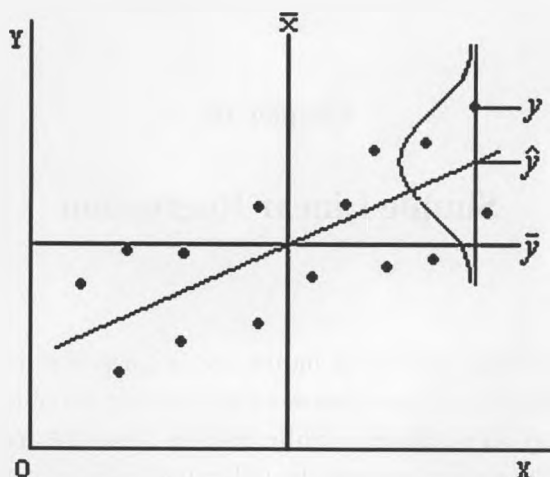


Fig. 10.1 A sketch map for regression.

polarization. Galton called this phenomenon “law of universal regression”. The term “regression” has been used in statistical sciences with thoroughly different meaning.

The relationship between two variables is always studied on the basis of paired sample data. If the values of X and Y are definitely one-to-one corresponding, their relationship can be described by an appropriate equation. However, due to the existence of variation among individuals and measurement error, one-to-one correspondence has never happened to the real data. If the paired sample data are plotted in a scatter diagram as showed in Fig. 10.1, although there is a linear tendency, the points do not exactly locate on a straight line. Based on the linear tendency, we may assume that corresponding to the values of X , the trace of the population mean of Y , denoted by $\mu_{y|x}$, can be located on a straight line. Such a linear relation between $\mu_{y|x}$ and X is called linear regression, which can be described with a linear regression equation as follows:

$$\mu_{y|x} = \alpha + \beta X, \quad (10.1)$$

where α is the intercept, which is the average level of Y when X takes value 0; β is the slope of the line, which is the increment of the average level of Y corresponding to increment of X by a unit. When $\beta > 0$, $\mu_{y|x}$ increases with X , the regression line ascends; when $\beta < 0$, $\mu_{y|x}$ decreases with X ,

the regression line descends; when $\beta = 0$, $\mu_{y|x}$ is independent of X , the regression line is parallel to the X -axis.

In general, the regression equation can only be obtained from sample, which is called sample regression equation or empirical regression equation. If we denote the sample estimate of $\mu_{y|x}$ as \hat{Y} , the sample regression equation can be expressed as

$$\hat{Y} = a + bX, \quad (10.2)$$

where \hat{Y} , a and b are the estimates of $\mu_{y|x}$, α and β .

10.1.2 Regression coefficient and its calculation

As showed in Eq. (10.2), once a and b are obtained from a sample, the sample regression equation is uniquely determined.

Viewing at the scatter diagram, to find a and b is equivalent to finding a straight line to best fit the points. In Fig. 10.1, the difference between the observed value y and the estimated value by the regression line \hat{y} is called a residual. The residuals could be positive as well as negative, of which the sum does not really reflect the discrepancy of the scatter points from the regression line. Therefore, in convention the sum of squared residuals are used to describe the fitness of the regression line; and one would like to find a straight line that minimizes the sum of squared residuals. This is so-called "principle of least squares". Under such a principle, it is easy to get the formulas for a and b by calculus as follows:

$$\begin{aligned} b &= \frac{l_{xy}}{l_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i) / n}{\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2 / n}, \end{aligned} \quad (10.3)$$

$$a = \bar{y} - b\bar{x}. \quad (10.4)$$

Putting the values of a and b into (10.2), one can get the linear regression equation. More than this, one may plot the regression line on the scatter diagram as an intuitive statistical description. Such a line must go through the point of (\bar{x}, \bar{y}) , and cross the vertical axis at a .

This is the part that has not been explained by the regression equation so it is called sum of squares for residuals or sum of squared deviations for errors.

(3) The regression makes the sum of squared deviations decline from SS_{Total} to SS_{Residual} , the contribution of the regression is

$$SS_{\text{Regression}} = SS_{\text{Total}} - SS_{\text{Residual}}. \quad (10.7a)$$

It can be proved that

$$SS_{\text{Regression}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = bl_{xy}. \quad (10.7b)$$

This is called sum of squares for regression. And its degree of freedom is

$$\nu_{\text{Regression}} = \nu_{\text{Total}} - \nu_{\text{Residual}} = 1. \quad (10.7c)$$

Obviously, the above steps showed a partition of the total sum of squared deviations and its degree of freedom:

$$SS_{\text{Total}} = SS_{\text{Regression}} + SS_{\text{Residual}}, \quad (10.8)$$

$$\nu_{\text{Total}} = \nu_{\text{Regression}} + \nu_{\text{Residual}}.$$

(4) To test H_0 : The contribution of regression is 0, a F -statistic is used,

$$F = \frac{SS_{\text{Regression}}/\nu_{\text{Regression}}}{SS_{\text{Residual}}/\nu_{\text{Residual}}} = \frac{MS_{\text{Regression}}}{MS_{\text{Residual}}}. \quad (10.9)$$

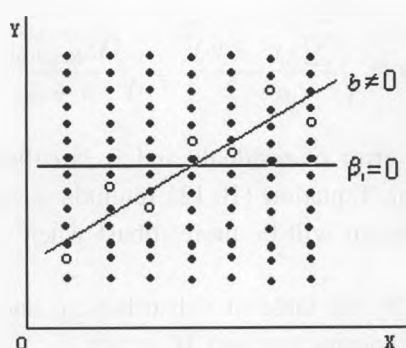
Here $MS_{\text{Regression}}$ and MS_{Residual} are called the mean sum of squares for regression and mean sum of squares for residuals respectively.

When H_0 is true, this F -statistic will follow a F -distribution with degrees of freedom $\nu_1 = 1$ and $\nu_2 = n - 2$. If the P -value is smaller than a pre-assigned α , then H_0 can be rejected at the level of α , and it is concluded that the regression is significant.

Finally the whole process can be summarized in a table of ANOVA (Table 10.2).

Table 10.2 Analysis of variance for regression.

Source	SS	DF	MS	F	P
Regression	$SS_{\text{Regression}}$	$\nu_{\text{Regression}} = 1$	$MS_{\text{Regression}} = \frac{SS_{\text{Regression}}}{1}$	$\frac{MS_{\text{Regression}}}{MS_{\text{Residual}}}$	
Residual	SS_{Residual}	$\nu_{\text{Residual}} = n - 2$	$MS_{\text{Residual}} = \frac{SS_{\text{Residual}}}{n - 2}$		
Total	SS_{Total}	$\nu_{\text{Total}} = n - 1$			

**Fig. 10.3** Possible sample when $\beta = 0$.

10.2.1.2 The *t*-test for regression coefficient

Figure 10.3 is a sketch showing $\beta = 0$, $\mu_{y|x}$ always stands on a horizontal line for any value of X . In such a case, $\mu_{y|x}$ does not depend on X so that we say the regression equation is not statistically significant.

However, even if $\beta = 0$ is true, it is not definitely impossible to have a sample like the blank circles in Fig. 10.3, then one may obtain a nonzero regression coefficient b . Therefore, when $\beta = 0$ is true, the difference between b and 0 is small, it is reasonable due to the existence of sampling error.

Contrarily, if the difference between b and 0 is fairly large, we may think that it is not very likely to have such a large b under the hypothesis of $\beta = 0$ and reject hypothesis H_0 . How large between b and 0 would we reject the hypothesis of $\beta = 0$? We need the following *t* test:

$$H_0 : \beta = 0, \quad H_1 : \beta \neq 0.$$

When H_0 is true, the statistic

$$t_b = \frac{b - 0}{s_b} \sim t \text{ dist.}, \quad (10.10)$$

$$\nu = n - 2, \quad (10.11)$$

where

$$s_b = \frac{s}{\sqrt{l_{xx}}}, \quad (10.12)$$

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{SS_{\text{Residual}}}{n - 2}}, \quad (10.13)$$

s is called standard error of residuals and s_b is called standard error of regression coefficient. Equation (10.12) reminds us that the estimate of the regression coefficient will be more robust when the values of X are spread out.

Same as before, by the table of t distribution, one can get the corresponding p value and decide to reject H_0 or not.

When both of the variables X and Y follow a bi-variable normal distribution, we can have correlation coefficient between Y and X as well as regression coefficient of Y on X . For the correlation coefficient ρ , we used to introduce a t test with a statistic t_r ; and now for the regression coefficient β , we also have a t test with a statistic t_b . It can be proved that these two t tests are equivalent, i.e. $t_b = t_r$.

Example 10.2 Work out a test for the regression in Example 10.1.

Solution (1) ANOVA The hypotheses to be tested are

H_0 : The contribution of the linear regression is 0,

H_1 : The contribution of the linear regression is not 0.

Calculate the total sum of squared deviations and its degree of freedom

$$SS_{\text{Total}} = \sum_{i=1}^n (y_i - \bar{y})^2 = 698.55,$$

$$\nu_{\text{Total}} = n - 1 = 20 - 1 = 19.$$

Calculate the sum of square for residuals, its degrees of freedom and the mean square for residuals

$$SS_{\text{Residual}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 94.92,$$

$$\nu_{\text{Residual}} = n - 2 = 20 - 2 = 18,$$

$$MS_{\text{Residual}} = \frac{SS_{\text{Residual}}}{\nu_{\text{Residual}}} = \frac{94.92}{18} = 5.27.$$

Calculate the sum of square for regression, its degree of freedom and the mean square for regression

$$SS_{\text{Regression}} = SS_{\text{Total}} - SS_{\text{Residual}} = 698.55 - 94.92 = 603.63,$$

$$\nu_{\text{Regression}} = \nu_{\text{Total}} - \nu_{\text{Residual}} = 1,$$

$$MS_{\text{Regression}} = \frac{SS_{\text{Regression}}}{\nu_{\text{Regression}}} = \frac{603.63}{1} = 603.63.$$

Calculate the ratio

$$F = \frac{MS_{\text{Regression}}}{MS_{\text{Residual}}} = \frac{603.63}{5.27} = 114.54.$$

Check the table for F -distribution with $\nu_1 = 1$, $\nu_2 = 18$, corresponding to $F = 114.54$, the P -value is less than 0.01.

Therefore, reject H_0 at the level of 0.01, and conclude that the regression of the son's height on the father's height is statistically significant.

Finally, summarize all the above results into a table of ANOVA (Table 10.3).

Table 10.3 Analysis of variance for regression in Example 10.1.

Source	<i>SS</i>	<i>DF</i>	<i>MS</i>	<i>F</i>	<i>P</i>
Regression	603.63	1	603.63	114.54	<0.01
Residual	94.92	18	5.27		
Total	698.55	19			

(2) t test The hypotheses to be tested are

$$H_0 : \beta = 0, \quad H_1 : \beta \neq 0.$$

By (10.13)

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{SS_{\text{Residual}}}{n - 2}} = \sqrt{\frac{94.92}{18}} = 2.2964.$$

By (10.12)

$$s_b = \frac{s}{\sqrt{l_{xx}}} = \frac{2.2964}{\sqrt{1859.2}} = 0.05326.$$

By (10.11)

$$t_b = \frac{b - 0}{s_b} = \frac{0.5698}{0.05326} = 10.7, \quad \nu = 20 - 2 = 18.$$

Check up the table for t distribution with $\nu = 18$, corresponding to $t_b = 10.68$, the P value is less than 0.001.

Therefore, reject H_0 at the level of 0.001, and also conclude that the regression of the son's height on the father's height is statistically significant.

Reviewing the test for correlation coefficient in Example 9.4, one will find that $t_b = t_r = 10.7$.

In addition, $t_b^2 = 10.7^2 = 114.54$, which is just equal to the F value in Table 10.3 obtained by the approach of ANOVA. This is not by chance. In theory, when the degree of freedom for numerator is equal to 1, the critical value of F distribution is equal to the square of that of t distribution,

$$F_{1,\nu} = t_\nu^2. \quad (10.14)$$

This shows the fact that the conclusion of t test for regression coefficient $\beta = 0$ is consistent with that of ANOVA for regression equation. A slight difference between these two approaches is that t test could be used for both one-side and two-side situations, but ANOVA for two-side only. However, the idea of ANOVA can be easily extended to the cases of nonlinear regression and multiple regression.

10.2.2 Determination coefficient

From Table 10.3,

$$SS_{\text{Regression}} = 603.63, \quad SS_{\text{Total}} = 698.55.$$

We have

$$\frac{SS_{\text{Regression}}}{SS_{\text{Total}}} = \frac{603.63}{698.55} = 0.8641.$$

At the same time, we have

$$r^2 = 0.9296^2 = 0.8641.$$

This is not by chance either. In fact,

$$\frac{SS_{\text{Regression}}}{SS_{\text{Total}}} = \frac{bl_{xy}}{l_{yy}} = \frac{l_{xy}^2}{l_{xx}l_{yy}} = r^2.$$

In general, the ratio between $SS_{\text{Regression}}$ and SS_{Total} is called determination coefficient or correlation index, denoted by R^2 ,

$$R^2 = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}}. \quad (10.15)$$

R^2 is a dimensionless quantity and $0 \leq R^2 \leq 1$. It reflects that the percentage of the total sum of squared deviations can be explained by the regression. Example 10.2 shows that 86.41% of the variation among the sons' heights can be explained by the information of their fathers' heights, while 13.59% of the variation cannot be explained yet.

In practice, it is suggested to report the value of determination coefficient after an analysis of regression to describe how good the regression is. Why? Here is a story: Once upon a time, one psychologist had a survey on the relationship between an index Y for liver function and a score for psychological status X through a group of patients with hepatitis B, resulting in a correlation coefficient $r = 0.2$ and a regression coefficient $b = 0.01$, and both were statistically significant. Although the psychologist had written in his report: "the index for liver function can be improved by psychological consultation", one year later he found that the effect of his psychological consultation was not as good as what he imagined. What is wrong? As a matter of fact, the determination coefficient here is only $0.2^2 = 0.04$. That

is, among the variation of the index for liver function, there is only 4% being determined by the score for psychological status.

10.2.3 Confidence interval for β

Based on the t test for regression coefficient, one can easily get a $(1 - \alpha)$ -confidence interval for the population regression coefficient β

$$b \pm t_{\alpha, v} \frac{s}{\sqrt{l_{xx}}}. \quad (10.16)$$

If this interval does not cover 0, we can also conclude that the population regression coefficient is not equal to 0 at a significant level of α . This is an alternative way to perform the hypothesis test.

Example 10.3 (Cont'd of Examples 10.1 and 10.2) Work out a confidence interval for the population regression coefficient in Example 10.1.

Solution From Examples 10.1 and 10.2, we have

$$b = 0.5698, \quad s = 2.2964, \quad l_{xx} = 1859.2.$$

By the table of t distribution we have $t_{0.05, 18} = 2.101$. By (10.16), the 95% confidence interval for the population regression coefficient β is

$$\begin{aligned} b \pm t_{\alpha, v} \frac{s}{\sqrt{l_{xx}}} &= 0.5698 \pm 2.101 \frac{2.2964}{\sqrt{1859.2}} \\ &= 0.5698 \pm 0.1121 = (0.48, 0.86). \end{aligned}$$

10.3 Applications of Linear Regression and the Pre-requisites

10.3.1 Two interval estimations

10.3.1.1 Confidence interval for $\mu_{y|x}$

Given $X = x_0$, by regression equation (10.2) one can get

$$\hat{Y}_0 = a + bx_0. \quad (10.17)$$

This is not the population mean $\mu_{y|x_0}$, only an estimate based on a sample. \hat{Y}_0 varies with the sample, of which the variation can be measured by a standard error

$$SE(\hat{Y}_0) = s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}. \quad (10.18)$$

From this formula, one can see:

- (1) The standard error of \hat{Y}_0 is proportion to s , the standard deviation of residual after regression;
- (2) When $x_0 = \bar{x}$, the corresponding $SE(\hat{Y}_0)$ reaches the minimum

$$SE(\hat{Y}_0) = \frac{s}{\sqrt{n}}.$$

- (3) The larger the distance $|x_0 - \bar{x}|$, the larger the standard error $SE(\hat{Y}_0)$ is. Based on (10.18), given $x_0 = \bar{x}$, we have the $(1 - \alpha)$ -confidence interval of $\mu_{y|x_0}$ as

$$\hat{Y}_0 \pm t_{\alpha, \nu} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}, \quad (10.19)$$

where $t_{\alpha, \nu}$ is the two-side critical value of t distribution with degree of freedom $\nu = n - 2$ corresponding to α . The belt formed by the two dot-curves in Fig. 10.4 is the geometric expression of (10.19). The belt takes the regression line as axis; it is the most narrow when $x_0 = \bar{x}$, and wider when $|x_0 - \bar{x}|$ is larger.

10.3.1.2 Prediction interval for Y

Given $X = x_0$, the corresponding individual value of Y_0 will vary around $\mu_{y|x_0}$. We have a confidence interval for $\mu_{y|x_0}$ so that the possible range of Y_0 will be wider than the interval. In fact, it can be proved that the standard deviation of Y_0 is

$$S(Y_0) \approx s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}. \quad (10.20)$$

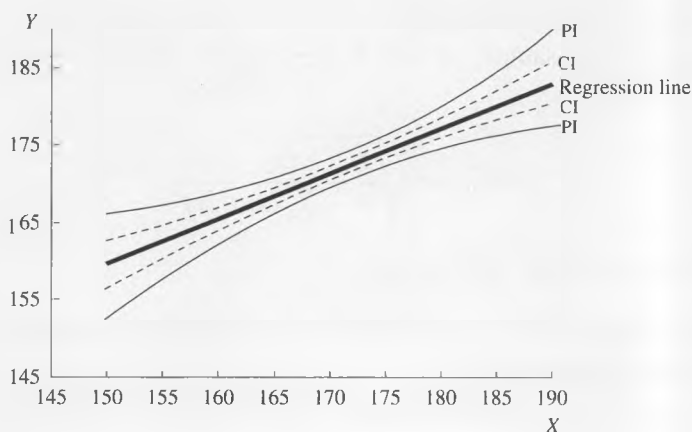


Fig. 10.4 The sketch map for the confidence interval of $\mu_{Y|X}$ (CI) and the prediction interval of Y (PI).

Therefore, given $X = x_0$, the $(1-\alpha)$ -prediction interval of the individual value of Y_0 is

$$\hat{Y}_0 \pm t_{\alpha, v} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}. \quad (10.21)$$

The belt formed by the two solid-curves in Fig. 10.4 is the geometric expression of (10.21). This belt also takes the regression line as axis, and it is most narrow when $x_0 = \bar{x}$, wider when $|x_0 - \bar{x}|$ is larger; but it surrounds the above-mentioned confidence interval of $\mu_{Y|x_0}$.

Example 10.4 Given the father's height $X = 165.8$ cm, apply the regression equation of son's height on father's height in Example 10.1 to estimate the 95% confidence interval for the average level of all the possible sons' heights and the 95% prediction interval for the specific son's height.

Solution From Examples 10.1 and 10.2, $a = 74.17$, $b = 0.5698$, $\bar{x} = 168.8$, $s = 2.2964$, $l_{xx} = 1859.2$, $t_{0.05, 18} = 2.101$. By (10.17), when $x_0 = 165.8$,

$$\hat{Y}_0 = 74.17 + 0.5698 \times 165.8 = 168.64.$$

By (10.19),

$$\begin{aligned}\hat{Y}_0 &\pm t_{\alpha, v} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \\ &= 168.64 \pm 2.101(2.2964) \sqrt{\frac{1}{20} + \frac{(165.8 - 168.8)^2}{1859.2}} \\ &= 168.64 \pm 1.1299 = (167.51, 169.77).\end{aligned}$$

By (10.21), the 95% prediction interval for the specific son's height is

$$\begin{aligned}\hat{Y}_0 &\pm t_{\alpha, v} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \\ &= 168.64 \pm 2.101(2.2964) \sqrt{1 + \frac{1}{20} + \frac{(165.8 - 168.8)^2}{1859.2}} \\ &= 168.64 \pm 4.9553 = (163.68, 173.59).\end{aligned}$$

The above example shows that given $X = x_0$, we can use the regression equation to estimate or predict the average level of Y_0 and the range of Y_0 . It has to be noticed that in general the estimation or prediction through the regression equation should be limited in the range of X that the regression equation comes from. In such a case, the estimation or prediction is called interpolation; otherwise, it is called extrapolation. If there is not enough reason to ensure that the relationship between the two variables keeps the same beyond the range that the regression equation comes from, extrapolation is not allowed.

10.3.2 Whether two data sets can be pooled for regression?

10.3.2.1 Are they parallel?

Assume that there are two samples

$$(x_{1i}, y_{1i}), i = 1, 2, \dots, n_1 \quad (x_{2i}, y_{2i}), i = 1, 2, \dots, n_2.$$

Work out two linear regressions respectively, we have

$$\hat{Y} = a_1 + b_1 X, \quad \hat{Y} = a_2 + b_2 X,$$

where (a_1, b_1) and (a_2, b_2) are the estimates of the parameters (α_1, β_1) and (α_2, β_2) of the two linear models respectively,

$$\mu_{y|x} = \alpha_1 + \beta_1 X, \quad \mu_{y|x} = \alpha_2 + \beta_2 X.$$

Sometimes we need to explore whether the two straight lines in the population are parallel, that is, to test

$$H_0 : \beta_1 = \beta_2, \quad H_1 : \beta_1 \neq \beta_2.$$

After the two regressions, denote the two sums of squares for residuals by $SS_{\text{Residual},1}$ and $SS_{\text{Residual},2}$; and

$$l_{xx,1} = \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2, \quad l_{xx,2} = \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2. \quad (10.22)$$

Since

$$s_{b1}^2 = \frac{SS_{\text{Residual},1}}{l_{xx,1}}, \quad s_{b2}^2 = \frac{SS_{\text{Residual},2}}{l_{xx,2}}$$

we have

$$s_{b1-b2} = \sqrt{\frac{SS_{\text{Residual},1}}{l_{xx,1}} + \frac{SS_{\text{Residual},2}}{l_{xx,2}}}, \quad (10.23)$$

where s_{b1-b2} is the standard deviation of $b_1 - b_2$.

It can be proved, when H_0 is true, that the statistic

$$t = \frac{b_1 - b_2}{s_{b1-b2}} \sim t \text{ dist.} \quad (10.24)$$

Given α , check the table of t distribution with degree of freedom $\nu = n_1 + n_2 - 4$, if $P \leq \alpha$, reject H_0 ; otherwise, do not reject, hence the two lines can be regarded as parallel to each other.

10.3.2.2 Can be pooled?

When the two lines are regarded as parallel, if the two lines can be further regarded as superposition, then the two sets of data can be pooled to have a unique regression equation.

The test hypotheses are

H_0 : The two lines are superposition in the population,

H_1 : The two lines are not superposition in the population.

On one hand, when H_0 is true, pooling the two data sets, one can have a regression equation $\hat{Y} = a + bx$ and denote the sum of squares for residuals with SS_{Residual} . On the other hand, when H_1 is true, working out two linear regressions respectively, we have two sums of squares for residuals $SS_{\text{Residual},1}$ and $SS_{\text{Residual},2}$.

Then calculate a statistic

$$F = \frac{[SS_{\text{Residual}} - (SS_{\text{Residual},1} + SS_{\text{Residual},2})]/\nu_1}{SS_{\text{Residual}}/(n_1 + n_2 - 2)}, \quad (10.25)$$

$$\nu_1 = 2, \nu_2 = n_1 + n_2 - 2.$$

Given α , check the table of F distribution with degrees of freedom ν_1 and ν_2 , if $P \leq \alpha$, reject H_0 ; otherwise, do not reject.

Example 10.5 Another school in northern China randomly selected 20 male students following the way of Example 9.2 to measure the heights (cm) of them and their fathers. The data are showed in Table 10.4. Work out a regression equation of Y on X with this data set. Suppose both Tables 10.1 and 10.4 come from a same collaboration project, can we pool the two data sets to have a unique regression equation?

Table 10.4 Heights (cm) of 20 pairs of father and son in a school of northern China.

No.	1	2	3	4	5	6	7	8	9	10
Father's height, X	154	159	166	168	170	172	174	176	178	179
Son's height, Y	158	168	167	173	168	175	170	171	179	180
No.	11	12	13	14	15	16	17	18	19	20
Father's height, X	179	180	181	181	182	182	185	186	188	192
Son's height, Y	174	181	175	181	175	183	176	183	185	180

Solution Based on Table 10.4, it is easy to get a regression equation

$$\hat{Y} = 67.63 + 0.6085X$$

and $s_{b2}^2 = 12.46$, $l_{xx,2} = 1726.8$ and $SS_{\text{Residual},2} = 224.28$. With a hypothesis test, it is concluded that this regression is statistically significant.

(1) Are they parallel?

To answer whether the two data sets can be pooled, we start from the test

$$H_0 : \beta_1 = \beta_2, \quad H_1 : \beta_1 \neq \beta_2.$$

Incorporating the results we have from Example 10.1: $b_1 = 0.5698$, $s_{b1}^2 = 5.27$, $l_{xx,1} = 1859.2$ and $SS_{\text{Residual},1} = 94.86$, we have the statistic

$$t = \frac{0.5698 - 0.6085}{\sqrt{\frac{5.27}{1859.2} + \frac{12.46}{1726.8}}} = -0.3860.$$

Check the table of t distribution with degree of freedom $\nu = n_1 + n_2 - 4 = 36$, we have $P > 0.7$. H_0 cannot be rejected so that the two lines can be regarded as parallel.

(2) Can be pooled?

Now let us test

H_0 : The two lines are superposition in the population,

H_1 : The two lines are not superposition in the population.

With the pooled data set, we get a regression equation

$$\hat{Y} = 70.58 + 0.5914X \quad (10.26)$$

and

$$SS_{\text{Residual}} = 320.72.$$

By (10.25), we have

$$F = \frac{[320.72 - (94.86 + 224.28)]/2}{320.72/38} = 0.0936.$$

Check the table of F distribution with degrees of freedom $\nu_1 = 2$ and $\nu_2 = 38$, we have $P > 0.10$, H_0 cannot be rejected so the two sets can be pooled and (10.26) is valid.

10.3.3 Linear regression through origin

In some specific applications, we are sure that Y must be 0 when $X = 0$. Then the model for regression is

$$\mu_{y|x} = \beta X. \quad (10.27)$$

The least square estimate of β is

$$b = \frac{\sum x_i y_i}{\sum x_i^2} \quad (10.28)$$

as if \bar{x} and \bar{y} in Eq. (10.3) were replaced by 0.

The linear regression through origin is a special case of general linear regression. The content in Secs. 10.2 and 10.3 are still valid as long as \bar{x} and \bar{y} in the formulas are replaced by 0 and the degrees of freedom change from $n - 2$ to $n - 1$.

10.4 On the Basic Assumptions and Analysis of Residuals

10.4.1 On the basic assumptions

The regression model (10.1) as well as (10.27) is subject to a kind of parametric models to describe the linear relationship between the two variables. If the basic assumptions are satisfied and able to get the support of the actual data, then the regression model can be used as guide for medical research. The basic assumptions for statistical inference related to this kind of models include the following four aspects:

- (1) There exists a linear tendency between the dependent variable Y and the independent variable X ("linear" for brief);
- (2) The individual observations are independent ("independent" for brief);
- (3) Given the value of X , the corresponding Y follows a normal distribution ("normal" for brief);
- (4) The variances of Y for different values of X are all equal, denoted by σ^2 ("equal variances" for brief).

As a summary, the basic assumptions could be expressed with "LINE" for brief. Failure to meet the basic assumptions might lead to a worse result, at least might affect the accuracy and precision of the estimates

and the validation of the P value. The assumption of linearity is essential that using a linear model to describe a curvilinear relationship is obviously inappropriate; the assumption of independency is also essential that one have to turn to more advanced approaches for dependent data; the violation to the assumptions of normal distribution and equal variance might not seriously affect the least square estimates though all the introduced formulas for statistical inference might not be valid. Once the assumptions (1), (3) and (4) are violated, it is worthwhile to try some transformations.

10.4.2 Analysis of residuals

In practice, one may use scatter diagram of residuals to observe whether the basic assumptions are met. Residual is defined as the corresponding Y minus the fitted value \hat{Y} predicted by regression model, i.e. $e = Y - \hat{Y}$. The examination of residual plots is a simple and effective method for detecting deficiencies in regression analysis.

The scatter plot of the residual e versus the fitted values \hat{Y} can be used to check the assumptions of linearity and equal variances. If the linear assumption does not hold, the plot may look like a curve. If the assumption of equal variances does not hold, the plot may look like the shape of a trumpet. The normal probability plot of the residual e or histogram of e can be used to check normality. The scatter plots of the residual against the predictor variable can be used to check dependency. If the plot shows a specific structure, the dependency may be violated. The Durbin-Watson statistics (DW) can also be used to check dependency. The value of DW is around 2 under this assumption; otherwise it will be close to 0 or 4.

Under the standard assumptions, the residual plot of residuals e versus the fitted values \hat{Y} should look like Fig. 10.5(a) which indicates that a linear model may be reasonable.

The scatter plot in Fig. 10.5(b) or (c) suggests a nonlinear model. One may build a curve model by nonlinear regression.

The scatter plot in Fig. 10.5(d), (e) or (f) shows that the residuals change with the fitted values indicating the assumption of equal variances is violated. One may estimate the model parameters by the way of weighted least squares method.

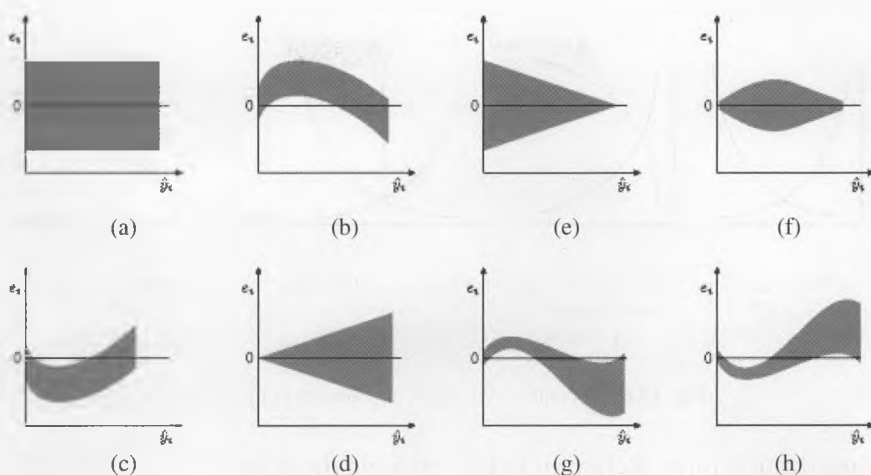


Fig. 10.5 Scatter plots of residuals versus the fitted values.

The scatter plot in Fig. 10.5(g) or (h) shows that the residuals change with the fitted values and has a curve shape which indicates linearity and equal variances are violated simultaneously.

If the assumptions of normality or equal variances dose not hold, one may make some transformations to variables such as logarithmic transformation, square-root transformation, and reciprocal transformation etc. It is worth noting that these transformations can be applied to independent variables but not the response variable.

10.5 Non-linear Regression

In medical practice and research, one may find that the relationship between two variables does not always appear in a linear pattern. When there exists a nonlinear relationship between the dependent variable Y and independent variable X , how to estimate the average level of Y corresponding to a given value of X ? This is the problem of nonlinear regression or curve fitting.

10.5.1 Through linear regression

As mentioned before, a scatter diagram is helpful to illustrate the possible relationship between two variables. Figure 10.6 demonstrates several curves often encountered in practice with the corresponding functions and their

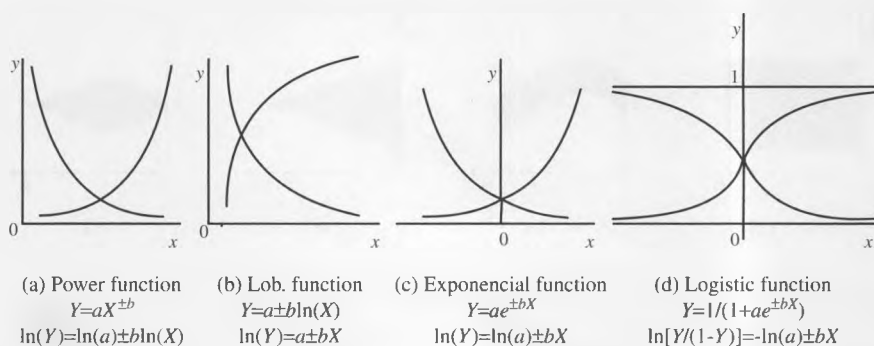


Fig. 10.6 Several curves often encountered in practice.

linearization form. Referring to Fig. 10.6, on the basis of scatter diagram, one may try various transformations to see which function may be chosen to fit the data.

Steps through linear regression:

- (1) Plot a scatter diagram for Y versus X ;
- (2) Attempt some appropriate transformations:

$$Y^* = f(Y), \quad X^* = g(X) \quad (10.29)$$

to lead the scatter diagram for Y^* versus X^* close to a straight line;

- (3) Perform a linear regression of Y^* on X^* ;
- (4) Substitute $Y^* = f(Y)$, $X^* = g(X)$ into the linear regression equation to obtain an equation for Y and X .

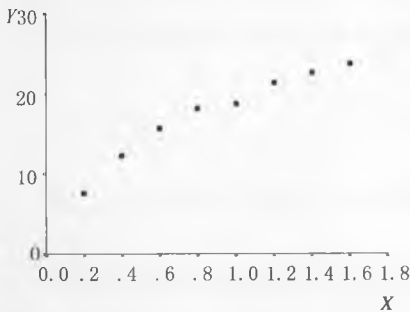
Facing a new situation, people prefer to solve the question by an old method first. To fit a curve round about a linear regression is just an example. It is feasible for some cases, but not all.

10.5.1.1 Linear regression after transformation of X

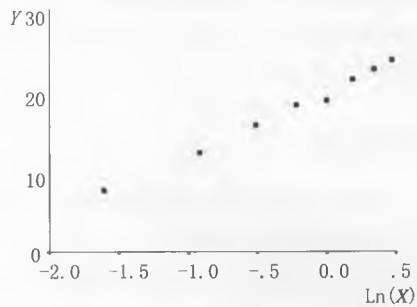
Example 10.6 The Department of Microbiology, Shanghai Medical University worked out an experiment of rocket electrophoresis to measure the rocket altitudes corresponding to given concentrations X of immunoglobulin A (IgA, $\mu\text{g/ml}$). The data are listed in Table 10.5. Try to fit a nonlinear regression equation for Y versus X .

Table 10.5 Data of an experiment of rocket electrophoresis.

IgA($\mu\text{g/ml}$) X	Rocket altitudes (mm) Y	$X' = \ln X$	\hat{Y}
0.2	7.6	-1.60944	7.22742
0.4	12.3	-0.91629	12.61907
0.6	15.7	-0.51083	15.77239
0.8	18.2	-0.22314	18.00972
1.0	18.7	0	18.74512
1.2	21.4	0.18232	21.16304
1.4	22.6	0.33647	22.36188
1.6	23.8	0.47000	23.40036



(a)



(b)

Fig. 10.7 Scatter diagram for Example 10.6. (a) $Y \sim X$, (b) $Y \sim \ln X$.**Solution**

- (1) Plot a scatter diagram for Y versus X . As showed in Fig. 10.7(a), has it looks like a curve.
- (2) Attempt the Logarithm function in Fig. 10.7(b). After a transformation of $X^* = \ln(X)$, a scatter diagram of Y versus X^* shows that it is linear.
- (3) Regression of Y on X^*

$$\hat{Y} = 19.980704588 + 7.639362628X^*. \quad (10.30)$$

The analysis of variance for linear regression shows a statistical significance with a determination coefficient $R^2 = 0.9922$.

(4) Substitute $X^* = \ln(X)$ into (10.30). We obtain

$$\hat{Y} = 19.980704588 + 7.639362628 \ln X. \quad (10.31)$$

It should be noted, in the above process, that we have not implied any transformation to Y so that among all possible equations with the shape of $Y = a \pm b \ln(X)$, (10.31) is the one minimizing the sum of squares of residuals $\sum (Y_i - \hat{Y}_i)^2$, hence it is the best answer following the principle of least squares.

10.5.1.2 Not recommended for any transformation to Y

Example 10.7 To study how the volume of sarcoma $Y(\text{cm}^3)$ increases with time X (day) for mice S78-3, a set of data has been collected and showed in Table 10.6. Try to fit a nonlinear function for the data through linear regression.

Solution

- (1) Plot a scatter diagram for Y versus X . As showed in Fig. 10.8(a), it looks like a curve.
- (2) Attempt the exponential function in Fig. 10.6(c). After a transformation of $Y^* = \ln(Y)$, a scatter diagram of Y^* versus X in Fig. 10.8(b) is linear.
- (3) Regression of Y^* on X

$$\hat{Y}^* = -4.27468874 + 0.156203483X. \quad (10.32)$$

The analysis of variance for linear regression shows a statistical significance with a determination coefficient $R^2 = 0.9517$. It implies that more than 95% of the variation of Y^* can be explained by X . We could say that the linear regression is satisfied.

- (4) Substitute $Y^* = \ln(Y)$ into (10.32)

$$\hat{Y} = 0.013916379e^{0.156203483X}. \quad (10.33)$$

Putting any value of X into Eq. (10.33), we can get the estimate of Y , denoted with \hat{Y}_1 . Column 3 of Table 10.6 gives all the values of \hat{Y}_1 corresponding to the observed value of X . One can see from Table 10.6 that regression of \hat{Y}^* on X seems satisfied by the way of least squares method but that of Y on X not. We will discuss it in Sec. 10.5.2.

Table 10.6 Data on volume of sarcoma $Y(\text{cm}^3)$ and time X (day) for mice S78-3.

Time (day) X	Volume of sarcoma (cm^3) Y	Linear regression				Nonlinear regression	
		Estimate of Y \hat{Y}_1	Estimate of residuals $e_1 = Y - \hat{Y}_1$	$Y^* = \ln Y$	\bar{Y}^*	Estimate of Y \hat{Y}_2	Estimate of residuals $e_2 = Y - \hat{Y}_2$
0	0.0042	0.0139	-0.0097	-5.4753	-4.2747	0.1609	-0.1567
6	0.0308	0.0355	-0.0047	-3.4809	-3.3375	0.2693	-0.2385
9	0.0614	0.0568	0.0046	-2.7901	-2.8689	0.3484	-0.2870
11	0.0744	0.0776	-0.0032	-2.5985	-2.5565	0.4136	-0.3392
13	0.1028	0.1060	-0.0032	-2.2750	-2.2440	0.4911	-0.3883
15	0.1516	0.1449	0.0067	-1.8863	-1.9316	0.5831	-0.4315
17	0.2101	0.1981	0.0120	-1.5601	-1.6192	0.6922	-0.4821
19	0.3390	0.2707	0.0683	-1.0817	-1.3068	0.8219	-0.4829
21	0.5201	0.3699	0.1502	-0.6538	-0.9944	0.9758	-0.4557
23	1.1020	0.5056	0.2567	-0.2714	-0.6820	1.1586	-0.3963
25	1.1020	0.6910	0.4110	0.0971	-0.3696	1.3756	-0.2736
27	1.5690	0.9444	0.6246	0.4504	-0.0572	1.6332	-0.0642
29	2.0214	1.2907	0.7307	0.7038	0.2552	1.9391	0.0823
31	2.7661	1.7641	1.0020	1.0174	0.5676	2.3023	0.4638
33	3.4289	2.4110	1.0179	1.2322	0.8800	2.7335	0.6954
35	4.1425	3.2951	0.8474	1.4213	1.1924	3.2454	0.8971
37	4.1593	4.5034	-0.3441	1.4254	1.5048	3.8532	0.3061
39	4.8590	6.1549	-1.2959	1.5808	1.8172	4.5749	0.2841
41	5.0037	10.4120	-5.4083	1.6102	2.1297	5.4317	-0.4280
43	6.3052	11.4967	-5.1915	1.8414	2.4421	6.4490	-0.1438
45	7.3461	15.7127	-10.3666	1.9942	2.7545	7.6568	-0.3107

10.5.2 Least squared estimate for nonlinear regression

In general, the nonlinear regression model can be expressed as

$$\mu_{Y|X} = f(\beta_1, \beta_2, \dots, \beta_p, X). \quad (10.34)$$

Here X refers to the independent variable(s); $\beta_1, \beta_2, \dots, \beta_p$ refer to the population regression coefficients; Y refers to the dependent variable, of which $\mu_{Y|X}$ is the population mean given X ; $f(\bullet)$ is a nonlinear function of $\beta_1, \beta_2, \dots, \beta_p$.

The basic assumptions of nonlinear regression model are all the same as those mentioned for the linear regression in Sec. 10.4.1 except the linearity. Therefore, we can still use the principle of least squares to estimate

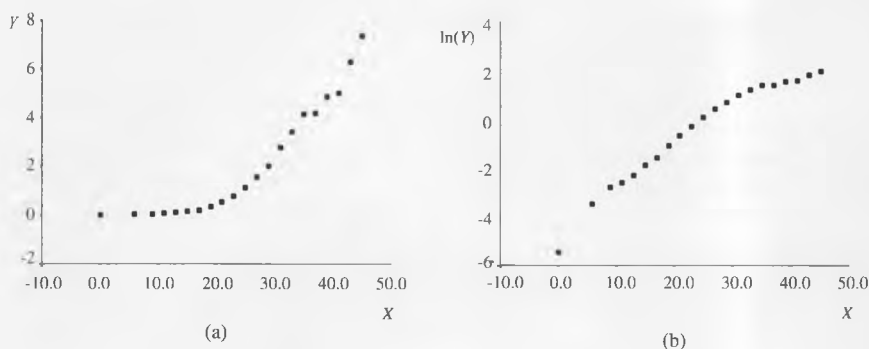


Fig. 10.8 Scatter diagram for Example 10.7. (a) $Y \sim X$, (b) $\ln Y \sim X$.

the sample regression coefficients, that is, to find suitable b_1, b_2, \dots, b_p to minimize the sum of squared residuals

$$Q = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

Since the expression of the nonlinear function $f(\bullet)$ might be complicated, we usually cannot get the explicit solution for the regression coefficients and hence numerical algorithms, such as the Newton–Raphson method, are used to get the estimates by iterations. The readers can easily find them from the commonly used statistical software.

In order to speed up the iteration, the most important thing is to assign a set of appropriate initial values to the unknown parameters. Referring to Sec. 10.4.1, if one can work out a linear regression without any transformation of Y , then the iteration for nonlinear regression is not necessary; if the scatter diagram can be turned to a linear pattern through a transformation of Y , then it is of no harm to get a set of initial values through a linear regression.

Example 10.8 Work out a nonlinear regression under the principle of least squares for the data in Example 10.7.

Solution Adopting the model of exponential function and using the Newton–Raphson method in statistical software SAS we can have the sample regression equation

$$\hat{Y} = 0.160892e^{0.085836X}. \quad (10.35)$$

Substituting the values of the observed X into this equation, the estimated values of Y are denoted with \hat{Y}_2 and listed in column 7 of Table 10.6; the estimated residuals are denoted by e_2 and listed in the last column 8 of Table 10.6.

Comparing the two sets of estimates \hat{Y}_2 and \hat{Y}_1 incorporating the observed values of Y , one will find that for $X \leq 21$, the estimate \hat{Y}_1 seems better than \hat{Y}_2 , and for $23 \leq X \leq 45$, \hat{Y}_2 is much better than \hat{Y}_1 . Why is that? It is mainly because the linear regression of Y^* on X is just responsible to Y^* , that is, the “least square” for Y^* does not guarantee the “least square” for Y .

From the success of Example 10.6 and the failure of Example 10.7, we know that: If the linear relationship can be performed by certain transformation of X only, then the result of linear regression might be satisfied; otherwise, the result of linear regression might be misleading if the linear relationship is performed by any transformation of Y .

How to compare the goodness-of-fit comprehensively for the nonlinear relationship?

10.5.3 Goodness-of-fit for nonlinear regression

Usually people use a determination coefficient R^2 to evaluate the result of regression for it is quite intuitive. The definition of it has been introduced in Chap. 8, that is, the proportion of the sum of squares contributed by regression to the total sum of squares of deviations, $SS_{\text{Regression}}/SS_{\text{Total}}$ or $1 - SS_{\text{Residual}}/SS_{\text{Total}}$.

However, this definition will fail to apply to nonlinear situations such as Example 10.7. In fact, based on column 2 of Table 10.2, we have the total sum of squares of deviation

$$SS_{\text{Total}} = \sum_i (Y_i - \bar{Y})^2 = 108.796321.$$

And based on columns 2 and 3, we have

$$SS_{\text{Residual}} = \sum_i (Y_i - \hat{Y}_{1i})^2 = 114.709638.$$

If it is the case,

$$R^2 = 1 - \frac{114.709638}{108.796321} = -0.0543522.$$

What a surprise! After regression, how can the sum of squares of residuals be greater than the total sum of squares of deviations before regression?

As we have seen by comparing columns 2 and 3 in Table 10.6, Eq. (10.33) is not as worse as nothing. The fallacy is caused by the definition of determination coefficient given in linear regression which is not appropriate for the situation of nonlinear regression.

We have mentioned in Chap. 9,

$$R^2 = \frac{[\sum (X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}.$$

Substituting the regression equation $\hat{Y}_i - \bar{Y} = b(X_i - \bar{X})$ into the above equation, we have

$$\begin{aligned} R^2 &= \frac{[\sum \frac{1}{b}(\hat{Y}_i - \bar{Y})(Y_i - \bar{Y})]^2}{[\sum \frac{1}{b^2}(\hat{Y}_i - \bar{Y})^2][\sum (Y_i - \bar{Y})^2]} \\ &= \frac{[\sum (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y})]^2}{\sum (\hat{Y}_i - \bar{Y})^2 \sum (Y_i - \bar{Y})^2} = r_{y, \hat{y}}^2. \end{aligned} \quad (10.36)$$

Thus, the determination coefficient R^2 is also a correlation coefficient between the observed Y and the estimated \hat{Y} in linear regression. Now we extend this to the situation of nonlinear regression as the definition of determination coefficient:

$$R^2 = [Cor(Y, \hat{Y})]^2. \quad (10.37)$$

For Example 10.6, it is a linear regression without transformation of Y so that the determination coefficient calculated according to Eq. (10.37) is still equal to 0.9922.

For Example 10.7, according to (10.37), the determination coefficient is the square of correlation coefficient between columns 2 and 3,

$$R_1^2 = [Cor(Y, \hat{Y}_1)]^2 = 0.9283.$$

For Example 10.8, according to (10.37), the determination coefficient is the square of correlation coefficient between columns 2 and 7,

$$R_2^2 = [\text{Cor}(Y, \hat{Y}_2)]^2 = 0.9857^2 = 0.9715.$$

Comparing R_1^2 and R_2^2 , one can see that the result of nonlinear regression is much better than that of linear regression after a transformation to Y . This is not surprising because the result of nonlinear regression is the best fitted one under the principle of least squares for Y rather than for any transformed Y .

In some old statistical literatures, the way of linear regression after transformations of X as well as Y to fit a curvature relationship were introduced, because it can be performed without the help of computer. Nowadays as the statistical software is quite popular, whenever the data cannot be well fitted by linear regression after transformation of X only, the least square estimation for a nonlinear model is recommended.

10.6 Computerized Experiments

Experiment 10.1 Linear regression Program 10.1 is used for linear regression analysis of Example 10.1.

Lines 02–14 of Program 10.1 generate a data set named A with two variables x and y . Lines 15 and 17 are to draw a scatter plot. Lines 18

Program 10.1 Linear regression analysis.

Line	Program	Line	Program
01	DATA A;	13	183 176 185 180
02	INPUT X Y @@;	14	;
03	CARDS;	15	PROC GPLOT;
04	150 159 153 157	16	PLOT Y*X;
05	155 163 158 166	17	RUN;
06	161 169 164 170	18	PROC CORR;
07	165 169 167 167	19	VAR X Y;
08	168 169 169 170	20	RUN;
09	170 173 171 170	21	PROC REG;
10	172 170 174 176	22	MODEL Y=X/R CLB CLI CLM DW;
11	175 178 177 174	23	SYMBOL I=RL;
12	178 173 181 178	24	RUN;

Program 10.2 The regression lines and their distribution after repeated sampling.

Line	Program	Line	Program
01	DATA A;	09	END;
02	DO I=1 TO 20;	10	PROC REG;
03	X=RANNOR(0)*2 + 170;	11	MODEL Y1=X;
04	ARRAY Y(10) Y1-Y10;	12	SYMBOL1 I=RL;
05	DO J=1 TO 10;	13	PROC GPLOT;
06	Y(J)=RANNOR(0)*2 + X*0.6 + 70;	14	PLOT (Y1 Y2 Y3 Y4 Y5 Y6 Y7
07	END;		Y8 Y9 Y10)*X = 1/OVERLAY;
08	OUTPUT;	15	RUN;

and 20 work for correlation analysis. Lines 21 and 22 work for regression. Line 23 is to plot a regression line.

Experiment 10.2 The distribution of sample regression lines Assume the father’s height X follows a normal distribution $N(170, 2^2)$, and the son’s height Y follows a normal distribution $N(70 + 0.6X, 2^2)$, that is, there is a “known” linear relationship between the population mean of the sons’ height and the father’s height, $\mu_{Y|X} = 70 + 0.6X$.

- (1) Randomly generate 20 pairs of the heights of fathers and sons, and work out a linear regression on such a data set;
- (2) Repeat (1) for ten times independently and obtain ten straight lines respectively;
- (3) Plot the ten lines on the same coordinate system;
- (4) Observe the distribution of these lines and discuss about the shape of the profile.

Lines 02–07 of Program 10.2 form a cycle to generate ten samples with 20 pairs of “observed values” for each. Lines 10 and 11 work for regression. Lines 12–15 plot regression lines of $Y1$ – $Y10$ versus X .

Experiment 10.3 Outcome of data-pool without consideration of confounding Let Z is a confounder affecting both the heights of father and son. Assume the father’s height follows a normal distribution $N(170 + 10Z, 4^2)$ and the son’s height follows a normal distribution $N(70 + 0.6(X - 10Z), 4^2)$.

- (1) Randomly generate three “samples” with 20 pairs of “father and son” corresponding to $Z = -1, 0, 1$ respectively;

Program 10.3 Observe the outcome of data-pool without consideration of confounding.

Line	Program	Line	Program
01	DATA A;	09	PROC PLOT;
02	DO I= -1 TO 1;	10	PLOT Y*X;
03	DO J=1 TO 20;	11	PROC REG;
04	X=RANNOR(0)*4 + 170 + 10*I;	12	MODEL Y=X;
05	Y=RANNOR(0)*4 + 0.6 + 70;	13	BY I;
06	OUTPUT;	14	PROC REG;
07	END;	15	MODEL Y=X;
08	END;	16	RUN;

- (2) Work out linear regression for the three samples respectively;
- (3) Pool the three data sets into one, and work out a linear regression;
- (4) Observe the difference between the results of (2) and (3), why?

Lines 02–08 in Program 10.3 generate three “samples” corresponding to $I = -1, 0$ and 1 . Lines 09 and 10 are to plot scatter diagram. Lines 11–13 work out linear regression for three “samples” respectively. Lines 14 and 15 work out regression and plot for the pooled data.

Experiment 10.4 Nonlinear regression Program 10.4 is used for multi-nomial logistic regression analysis of Example 10.6.

Lines 02–10 of Program 10.1 generate a data set named A with two variables X and Y . Lines 13 and 18 work for nonlinear regression of Y on X .

Program 10.4 Nonlinear regression by least squares method.

Line	Program	Line	Program
01	DATA A;	10	41 5.0037 43 6.3052 45 7.3461
02	INPUT X Y@@;	11	;
03	CARDS;	12	RUN;
04	0 0.0042 6 0.0308	13	PROC NLIN;
05	11 0.0744 13 0.1028 15 0.1516	14	PARMS B0=0.013916379
06	17 0.2101 19 0.339 21 0.5201		B1=0.156203483;
07	23 0.7623 25 1.102 27 1.569	15	MODEL Y=B0*EXP(B1*X);
08	29 2.0214 31 2.7661 33 3.4289	16	OUTPUT OUT=B P=YHAT R=R;
09	35 4.1425 37 4.1593 39 4.859	17	RUN;

Experiment 10.5 Comparison linear regression after transformation

of Y and nonlinear regression Generate a set of data, of which the mean follows an exponential function $\mu_{Y|X} = e^{-0.5X}$, and the residual follows a normal distribution, then fit an exponential function by linear regression after transformation of Y and nonlinear regression respectively and finally compare the results.

The detailed steps are the follows:

- (1) Calculate the values of $e^{-0.5X}$ for $X = 0, 1, 2, 3, 4$;
- (2) For each value of X , generate 20 values of Y from the normal population $N(e^{-0.5X}, 0.1^2)$;
- (3) Plot a scatter diagram for Y versus X and observe the distribution of Y for different values of X ;
- (4) Plot another scatter diagram for $\ln Y$ versus X and observe the distribution of $\ln Y$ for different values of X ;
- (5) Compare (3) and (4);
- (6) Estimate the parameters α and β for the model $\mu_{Y|X} = \alpha e^{-\beta X}$ through a linear regression of $\ln Y$ on X ;
- (7) Estimate the parameters α and β for the model $\mu_{Y|X} = \alpha e^{-\beta X}$ through a nonlinear regression of Y on X ;
- (8) Compare the results of (6) and (7) to check which is closer to the initial equation $\mu_{Y|X} = e^{-0.5X}$.

Program 10.5 Linear regression after transformation of Y and nonlinear regression.

Line	Program	Line	Program
01	DATA A;	14	PLOT (EY Y)*X/OVERLAY;
02	DO X=0 TO 4 BY 1;	15	PROC GPLOT;
03	DO I=1 TO 20 BY 1;	16	PLOT (LEY LY)*X/OVERLAY;
04	Y=EXP(-0.5*X);	17	PROC NLIN;
05	EY=EXP(-0.5*X) +RANNOR(0)*0.1;	18	PARMS B0=0.9 TO 1.1 BY 0.05
06	LEY=LOG(EY);		B1=0.4 TO 0.6 BY 0.01;
07	LY=LOG(Y);	19	MODEL EY=B0*EXP(-B1*X);
08	OUTPUT;	20	DER.B0=EXP(-B1*X);
09	END;END;	21	DER.B1=-B0*X*EXP(-B1*X);
10	PROC PRINT;	22	PROC REG;
11	VAR EY LEY;	23	MODEL LEY=X;
12	BY X;	24	RUN;
13	PROC GPLOT;		

Lines 02–09 of Program 10.5 form two cycles to generate the “observed values” of Y from the normal population $N(e^{-0.5X}, 0.1^2)$ for $X = 0, 1, 2, 3, 4$, denoted with EY; lines 06 and 07 take logarithms for Y and EY, denoted with LY and LEY. Lines 10 to 12 print the generated data. Lines 13–16 plot the scatter diagrams of Y , EY, LY and LEY versus X respectively. Lines 17–21 estimate the parameters by nonlinear regression based on X and EY; line 18 defines the range of the parameter to be estimated; lines 20 and 21 define the partial derivatives for the two parameters. Lines 22 and 23 estimate the parameters by linear regression based on X and LEY.

10.7 Practice and Experiments

1. True or false? Why?

- (1) The linear regression equation does not change even though the units of X and Y are changed.
- (2) “Least square” means that the sum of the differences between the observed values and the values calculated with the regression equation is minimized.
- (3) For the same sample, when $b = 0$, there must be $r = 0$; and when $b = 1$, there must be $r = 1$.
- (4) The linear regression equation requires that both X and Y should follow normal distribution respectively.
- (5) If two regression lines are of superposition in the population, then the two correlation coefficients must be equal.
- (6) The larger the determination coefficient, the better the regression is.
- (7) If the regression equation of high school student’s height Y (m) on their age X (year) is $\hat{Y} = 0.5 + 0.06X$, then the average height of newborn is 0.5 because $\hat{Y} = 0.5$ when $X = 0$.
- (8) In the hypothesis test for regression coefficient $H_0 : \beta = 0$, the smaller the P value, the larger the $|\beta|$ is.
- (9) If one can predict Y by X , there must be causal relationship between X and Y .
- (10) It is easy to get a $P < 0.05$ in the test for regression when n is big enough so that as long as the sample size increases we can always find a variable severely affected by another.

2. It is reported that the correlation coefficient between X and Y is $r = 0.90$, the sample means and standard deviations of them are $\bar{x} = 50$, $\bar{y} = 100$, $s_X = s_Y = 10$. Can you quickly write the regression equation based on the information? When both X and Y are random variables, can you find out the relationship between correlation coefficient and regression coefficient from the formulas for r and b ?

3. Given a random sample of X and Y , (x_i, y_i) , $i = 1, 2, \dots, n$, denote the sample means and standard deviations of them with \bar{x} , \bar{y} , s_X , s_Y , and the standardized variables with X^* and Y^* . It can be proved that the regression coefficient of the regression of Y^* on X^* must be equal to the correlation coefficient between X and Y . Observe this rule from the data of Example 10.1.

4. Based on the data of problem 5 in Chap. 9, answer the following questions:

- (1) Work out a regression equation for the score of mathematics (X) on the intelligent quotient (Z); then calculate the correlation coefficient between the residual $X - \hat{X}$ and Z . What does the result mean?
- (2) Work out a regression equation for the score of literature (Y) on the intelligent quotient (Z); then calculate the correlation coefficient between the residual $Y - \hat{Y}$ and $X - \hat{X}$. What does the result mean?

5. Work out the formula of the 95% confidence interval for the regression coefficient β in Example 10.5; given a father's height $x_0 = 175$ cm, estimate the 95% confidence interval for the average level of all the possible sons' heights and the 95% prediction interval for the specific son's height.

6. If every value of Y in Example 10.5 is added by 10 cm, what change will happen to the regression equation? After adding 10 cm to every value of Y , whether such a new data set can be pooled with the data in Example 9.2 to have a unique regression equation?

7. Give an example of testing the parallelism of two regression lines from your field of subject matter. What is the real meaning of "parallelism" in this example? What is the real meaning of "not superposition" in this example?

8. Summarize the methods for judging if the regression equation is significant or not?

Table 10.7 A set of data on the attack rate of measles vs. time.

Time X	1	2	3	4	5	6
Attack rate (%) Y	34.3	65.5	76.8	85.2	90.3	94.1

9. Summarize the points to be noticed in linear regression analysis.
10. Briefly summarize the relationship and distinguish between linear regression and linear correlation.
11. Yibei Ling (1987) reported a set of data from a project on the attack rate of measles versus time as showed in Table 10.7. Use two models (exponential and logarithm) to fit the data by two procedures (linear regression after transformation and nonlinear regression) respectively.
 - (1) For each model, is there any difference between the results obtained by the two procedures?
 - (2) For each procedure, which model is more appropriate to this data set? (From Peihuan Jin, China Health Statistics, 1987, 4(4).)

12. Table 10.8 is the census data of the US population (excluding Hawaii and Alaska). Peal, Reed and Kish had fitted a logistic model for the data from 1790 to 1940 as follows:

$$\hat{W} = \frac{184}{1 + 66.69(10^{-0.1398X})}$$

From the column for residual in Table 10.8, one can see that the fitness is really good. However, when it was used to predict the populations of 1950 and 1960, the residuals were terribly large. Why? What do you think about?

13. A survey on 2557 primary and high school students in a city for their HbsAg infectious situation resulted in a set of data showed in Table 10.9. Fit an exponential model for this data set by both linear regression after transformation and nonlinear regression respectively, and compare their goodness-of-fit. (From Gen Lu, China Health Statistics, 1993, 10(1)).
14. A survey on the costs for hospitalization per inpatient in a hospital during the period from 1977 to 1989 resulted in a data set given in Table 10.10.

Table 10.8 The populations of the US (not include Hawaii and Alaska) and the estimates.

Population (million)				Population (million)			
Year	Actual \bar{W}		Residual $W - \bar{W}$	Year	Estimate \hat{W}		Residual $W - \hat{W}$
	Actual	Estimate			Actual	Estimate	
1(1790)	3.9	3.7	+0.2	10(1880)	50.2	50.2	0.0
2(1800)	5.3	5.1	+0.2	11(1890)	62.9	62.8	+0.1
3(1810)	7.2	7.0	+0.2	12(1900)	76.0	76.7	-0.7
4(1820)	9.6	9.5	+0.1	13(1910)	92.0	91.4	+0.6
5(1830)	12.9	12.8	+0.1	14(1920)	105.7	106.1	-0.4
6(1840)	17.1	17.3	-0.2	15(1930)	122.8	120.1	+2.7
7(1850)	23.2	23.0	+0.2	16(1940)	131.4	132.8	-1.4
8(1860)	31.4	30.3	+1.1	17(1950)	150.7	143.7	+7.0
9(1870)	38.6	39.3	-0.7	18(1960)	178.5	152.9	+25.6

Table 10.9 Infection rates of primary and high school students in a city.

Grade X	1	2	3	4	5	6
Infection rate (%) Y	3.57	4.14	3.25	4.14	5.44	3.82
Grade X	7	8	9	10	11	
Infection rate (%) Y	4.19	5.12	6.15	5.93	6.77	

Table 10.10 The costs for hospitalization per inpatient in a hospital (1977–1989).

Year	1(1977)	2(1978)	3(1979)	4(1980)	5(1981)
Cost per inpatient (Yuan)	36.32	37.15	38.14	41.20	44.39
Year	6(1982)	7(1983)	8(1984)	9(1985)	10(1986)
Cost per inpatient (Yuan)	43.48	64.79	127.01	177.68	221.30
Year	11(1987)	12(1988)	13(1989)		
Cost per inpatient (Yuan)	296.24	477.14	634.49		

Fit the data by both nonlinear regression and segmental linear regression respectively; compare the results and discuss the potential problem of prediction. (From Lanhua Chen and Binhui Wang, China Health Statistics, 1992, 9(2)).

(1st edn. Jiqian Fang; 2nd edn. Jinxin Zhang, Jiqian Fang)

Chapter 11

Statistical Principles for Design of Interventional Study

In the design of medical studies, researchers develop protocols based on professional knowledge and statistical theory to investigate specific medical problems. A good research protocol can lead to reliable results and conclusions with limited time and resources. According to whether patients receive interventions from the researchers, medical studies can be classified into interventional study and observational study.

Interventional study, in which the investigator determines to give specific treatments to some objects and control treatments to others, is widely used in pre-clinical medicine, clinical medicine, preventive medicine and so forth. The results reflect the effects both from the controllable treatment factors and some uncontrollable non-treatment factors. For example, in a clinical trial, curative effect of a drug is not only affected by the drug itself, but also affected by the route and time of administration, patients' physical condition, even psychological conditions of patients and doctors, etc. Since many non-treatment factors, of which some can be controlled but others cannot, affect the outcome, we must pay enough attention to a good research design before hand. Research design is a comprehensive plan with meticulous care for the allocation of treatments, determination of the measurement and analysis of data, to guarantee the balance of non-treatment factors between contrast groups so that the outcome may have better comparability and well-controlled error to obtain the reliable conclusion from not too large sample.

Observational study is also known as survey study. It is a survey of specific populations without interventions. Through objective observation and recording, researchers can describe the results and analyze the relationships between factors and outcome(s). Comparing with interventional study, observational study proceeds in a most “natural-like” status, with all factors and levels of factors of a patient being decided naturally, not by researchers or randomly. According to different study design, the types of observational study include cross-sectional study, cohort study and case-control study.

In this chapter, some basic concepts and rules in research design, especially, in design of interventional study will be introduced. As the most commonly used methods in clinical trial, randomized controlled trial will also be introduced in this chapter.

11.1 The Essential Concepts of Design

11.1.1 *Three elements of a medical research*

There are three elements in a medical research, namely treatment, subject and effect. For instance, to evaluate the effect of a hypotensor, using the hypotensor or not is the treatment, the hypertensives are the subjects, and the decline of blood pressure is the effect.

11.1.1.1 *Treatment*

According to research purpose, the researcher wants to observe the direct or indirect effects of some factors, which act on the individual unit. These researcher-defined factors are named as treatment.

In medical research, besides the effect of treatment factors, some non-treatment factors can also affect the outcome. For example, the condition of making substrate, the location and duration of placed substrate can confound the effects of the treatment factor. In order to determine the treatment effect, we should try our best to find the important non-treatment factors that influence the research result so that we can remove or weaken their effects.

Treatment factors should be kept standardized, in other words, the treatment should be consistent during the whole study.

11.1.1.2 *Subject*

Based on different purposes, the research subject can be a person or animal and even a certain organ, serum, cell, etc. If the research subject is a patient, one may call the research as clinical trial, otherwise experiment.

The valid range or condition of research subject must be predefined explicitly so as to ensure the homogeneity. All research subjects satisfying those conditions are called a population, and the subjects included in the study are called samples. The conclusion can be generalized if the samples are valid and the characteristics of the population are met.

11.1.1.3 *Effect*

The “effect” is the outcome after the treatment and often expressed by a few relevant indices (variables). The selecting matter on indices and measurements have great influence on the validity of a research and always relies on the subject matter knowledge. Here we just mention some essential criteria: valid, precise and sensitive.

11.1.2 *Errors and their characteristics*

11.1.2.1 *Random error and its characteristics*

Random error results from a great deal of exiguous and incidental errors that are hard to control one by one. Although we cannot estimate the random error by only one observation, under a number of repetitions, the random error presents certain regulation, and generally obeyed the normal distribution with zero as the mean. One of the important missions of the research is to help researches make use of the regulation of random error and reveal the objective by statistical analysis.

11.1.2.2 *Non-random error and its characteristics*

Non-random error, also named as bias or systematic error, occurs due to the conditions other than the treatment and makes the result deviated from the truth systematically. Different from random error, without well prevention from the systematic error, one cannot simply make use of the statistical analysis to make any inference. Another important mission of research design is to lower or eliminate the bias within the result.

Bias can come from every stage of the research process. According to its sources, bias can be classified into the following types:

(1) Selection bias It occurs at the early stage of research. Due to the inappropriate choice of the research subjects, the groups lose comparability and the results contain bias. For example, in a clinical trial, when the physicians do not restrictively comply with the inclusion and exclusion criteria of protocol, say, assigning patients to different groups according to the patients' wills may produce bias to the results.

(2) Information bias It is also called metrical bias and occurs at the moment the subjects are measured. This bias happens when the measuring instruments have not yet been calibrated, the criterion of the operation is not standardized or subjective preference exists during the analysis of results. It also includes the gross error resulting from falsely recorded data.

(3) Confounding bias Confounding occurs when the non-treatment factors affect the results unfairly to the compared groups. As we know, the sequel of a disease not only attributes to the treatment function of a medicine, but also relates to its natural development, accessory treatment and patient's constitution. If ones only notice the association between medicine and disease, and neglect the balance of other factors among the compared groups, the confounding bias would lead to a wrong conclusion.

In summary, the non-random error or the bias may take place in each stage of medical research and affects the conclusion to some extent. Once the bias occurs, we can hardly rectify or make up after that. Therefore, the researcher must carry on the study cautiously at each stage to avoid bias or lower down its influence on the conclusion.

11.1.3 *The statistical principles for research design*

In order to control the random error of a study, avoid or reduce the non-random error preferably, meanwhile obtain more reliable information by observing less subjects, we should fulfill the following three statistical principles during the design stage.

11.1.3.1 *Control*

Discrimination is always based on a rational comparison. A "nice" control group must be established within an interventional study in order to manifest the effect of treatment (see Fig. 11.1).

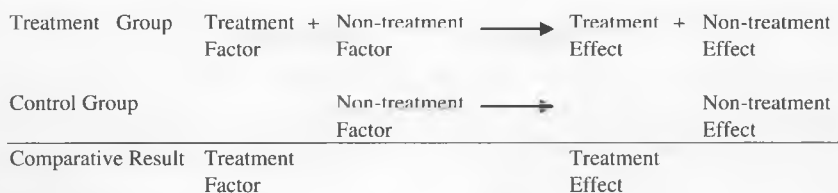


Fig. 11.1 A sketch of the function of control group.

The balance between control and treatment groups is the premise that ensures the right manifestation of effects of the treatment. Balance means that among the compared groups, only the treatment factors are different, the distribution of those important and controllable non-treatment factors should keep consistent as long as possible. For example, the research subjects within different contrast groups should keep consistent on the distributions of gender, age and health condition at the beginning. For a clinical trail, we should consider for every patient the seriousness and the course of disease and other treatments that one had accepted etc.

In medical research, the constitution of control group must meet three conditions:

- (1) **Equity** Except for treatment factors, the control group must have the same non-treatment factors as the treatment groups do.
- (2) **Synchronization** Once the control and treatment groups set up, we should guarantee that the whole research process for both groups take place in the same time and space.
- (3) **Specificity** The control group is established exclusively for the relevant treatment groups on study. We should not use other's results or fall back on literatures as the control for the current study.

These three conditions guarantee the balance between control and treatment groups, thus the function of control group can be completely implemented. After establishing the control group, we should compare the base line conditions of all compared groups to examine their balance.

The principle of "control" is also important to observational studies because comparison is the soul of any study and comparability should always be concerned.

11.1.3.2 *Randomization*

“Randomization” is a kind of statistical operation in order to enhance the consistency of distributions of many uncontrolled non-treatment factors among all the compared groups. Randomization should be carried out during the process of sampling, grouping and executing:

- (1) Random sampling All the subjects who meet the fixed criterion have the same opportunity to be included, in other words, any individual within the population has the same probability to be taken into the sample;
- (2) Random allocation Any included subject has the same probability to be assigned to each of the compared groups;
- (3) Random treated orders In some design, if the subjects do not receive the treatment at the same time, then any subject has the same probability to receive the treatment early or later.

Here, random sampling guarantees to get representative sample and makes the experiment conclusion holds for the population; random grouping enhances the balance and comparability among compared groups; random experiment orders eliminate the influence from the order of receiving the treatment. The method for randomization can be found in Sec. 11.3.

The principle of “randomization” is also important to observational studies except that of “random allocation” because the researcher in observational study cannot allocate the treatment to the subject, but observe what occurs in the “real world”.

11.1.3.3 *Replication*

“Replication” means that we take a number of observations under the same experimental condition so as to improve the reliability and validity of an experiment result. On the broad sense, replication implies:

- (1) The replication of whole experiment.
- (2) Carry out the experiment with more experiment units.
- (3) The repeated observations on the same experiment unit.

Here, (1) makes the experiment replicable and improves its reliability; any result not being replicable cannot be scientific. (2) avoids taking specific as universal or taking coincidence as necessity; through certain amount of

repetition we can make the conclusion reliable. (3) aims to enhance the precision of the measurement, for example, the blood pressure is generally measured three times, and the average is taken as the final observation value.

From probability theory, we know that the more repetition of experiment unit, the closer between parameter and sample statistics, say, sample frequency or sample mean etc. However, too many observations will waste resource, and even make researchers fail to control the experimental condition efficiently, thus lower the reliability of experiment result. From the computation of standard error, we notice that if the sample size extends up to 100, theoretically the standard error only reduces to the 1/10 of its original value. Obviously, the loss cannot compensate for the gain. One purpose of statistical design is to estimate adequate sample size to make the statistical conclusion reliable and avoid unnecessary waste.

In addition to previously mentioned conditions, the determination of sample size involves many other factors. For instance, the sample size needed for quantitative index is less than that for qualitative index; under the same test level and test power, the sample size needed for one-side test is less than that for two-side test; if the numbers of individuals in the contrast groups are the same, the total sample size needed is less.

R. A. Fisher firstly suggested the must-be-fulfilled three principles above in 1935. Since then, many other design methods or techniques have been put forward by statisticians, but the basic ideas are all comply with the principle of control, randomization and replication, just the appearances may vary.

11.2 Statistical Principle in Clinical Trials

Clinical trial is a kind of prospective researches aims at evaluating a clinical intervention by comparing effect of this intervention and that of the control treatment. Different from animal experiment, in clinical trial human patient is the research object. Researchers cannot absolutely dominate the patients' actions in a clinical trial, they can only request the patients to avoid some actions disturbing experiment results, therefore, the patients' compliance and ethnic problem must be considered. Once the new drug has been confirmed harmful to patients, the trial must be terminated. When there is a drug of which the curative effect for the disease under study is validated, using placebo contrast is inappropriate. When emergency situation happens, the

patient must receive additional treatment. Thereby in a clinical trial, much more problems exist and need more restrict and specific request in design.

Clinical trail for a new drug is divided into four phases: Phase I clinical trial includes the initial introduction of an investigational new drug into humans, preliminary clinical pharmacology and safety evaluation, which is designed to observe the pharmacokinetics and pharmacological actions of the drug in humans, provide basic information for further prescription scheme; phase II clinical trail is a randomized, blinded, controlled study aiming to obtain some preliminary data on the effectiveness and safety of the drug for a particular indication and recommend clinical dosage; phase III clinical trial is an extended multi-center experiment with randomization and control, which makes further evaluation on effectiveness and safety; phase IV clinical trail aims to monitor the curative effects and inverse effects, especially rare inverse effects, under the extensive use after the new drug came into the market. In this chapter, we mainly discuss the statistical requirements of phases II and III clinical trails.

Example 11.1 The physicians notice that after the acute myocardial infarction (AMI) happens, infarct expansion and left ventricular remodeling often lead to left ventricular augmentation and cardiac dysfunction. In order to evaluate the effect of long-term Captopril after AMI, one needs to design a clinical trial and the following questions should be considered in the design.

11.2.1 *The choice of the outcome variables*

11.2.1.1 *Primary outcome variable and secondary outcome variable*

The primary outcome variable, also called target variable, can provide reliable evidence on the purpose of the clinical trial. Generally there is only one primary outcome variable that should be objective, easy to measure and widely accepted in related research field. The secondary variable refers to the additional variable related to the main purpose of the trial, or variable related to the secondary purpose. Both kinds of variables should be explicitly defined and the reasons being chosen explained in the design protocol. The determination of sample size, evaluation of the effectiveness and safety should be based on the primary outcome variable.

11.2.1.2 *Compound outcome variable*

When there are many variables related to the main purpose of a clinical trial, it is difficult to choose one single primary variable. Here we can choose some computation techniques in advance (say, sum or weighted sum, etc.), or incorporate many variables into one variable. For instance, take “stroke or hospitalization due to cardiac disease” as a compound outcome variable.

11.2.1.3 *Global assessment outcome variable*

This sort of variables is synthesized from some objective variables and researcher’s total impression on the patients’ condition and is usually a scale of ordered categorical ratings. Therefore, if it is indeed needed, one should explain explicitly in the protocol that it is relevant to the main purpose for the trial, it has adequate and reliable reason to be chosen and there exist unambiguous rules to judge the grade. The global assessment variable with better objectivity should be considered alone as one of the primary variables.

In Example 11.1, the main purpose is to appraise benefits from Captopril for the pumping capacity of the left ventricle; hence the primary outcome variable should be the capacity of left ventricle and ejection fraction measured by Doppler echocardiography. The secondary outcome variable could be the left ventricular filling rate. There exist more than one variable to appraise the capacity and filling rate of left ventricle, when inconsistent results come forth, it is difficult to get a unique conclusion. Therefore one can aggregate several variables into one compound variable for statistical analysis. In this example, the main purpose of the research is to evaluate the protective function of Captopril to the heart so that an ordered categorical variable reflecting heart function would be a global assessment variable.

11.2.2 *The choice of the control groups*

In clinical trial, the only difference between control and treatment groups is that the patients in treatment groups accept the new drugs, but those in control groups accept the contrast drugs (or placebo).

It is requested that both treatment and control groups come from the same population. An ideal setup would be a similar baseline at the beginning

and consistent conditions along the progress of the trial for both groups. In clinical trial, the usual setup of control groups involves the following three types:

11.2.2.1 *Placebo control*

The placebo is a kind of fake drug. Its attribute, such as the form of appearance, size, color, weight, smell and taste etc., all keep unanimity with that of experimental drug as much as possible, but does not contain any active ingredient of the experimental drug. The purpose of placebo control lies on removing the bias evoked from the psychological factors of the physicians, the patients and other participants in the trial, and isolates the real effects and inverse effects caused by the drug. The placebo can be used in parallel or crossover setup.

11.2.2.2 *Active control*

It refers to an effective drug that has been authorized to go into the market. It must be legal, safe, and generally accepted as most effective for the disease under investigation. It can be used in parallel or crossover setup.

11.2.2.3 *Dose-response control*

In this setup, researches devise the dosage of the experimental drug into several segments, and the patients are arranged into different dosage groups at random. There, the placebo contrast, namely the zero-dose group, can be involved or not. This setup is mainly used for illustrating the relationship between the dosage and the curative effect or the inverse effect, or just the former. The kind of control redounds to the approval of the dosage.

In Example 11.1, the patients suffering from AMI are not suitable to take the placebo control, therefore they should take the routine treatment as an active control, say, the conventional clot-dissolving drug, aspirin etc. The treatment group should take the routine treatment plus Captopril.

According to the practical situation, more than one control groups could be established in a clinical trial. When placebo, active control and the trial drug groups present at the same time, it is called three-arm study. In a placebo control experiment, according to medical ethics considerations, a standard

drug is simultaneously added to treatment group and placebo group, this setup is called placebo-standard study.

11.2.3 *The important skill for avoiding bias — blinding method*

The blinding method aims at prevent participants in the clinical trial, including the sponsor, investigator, medical personnel, monitor, data manager and statistical analyst, from knowing which patient accepts what treatment, thus avoid the impact of their subjective judgments to the result. There are two states of blinding, namely double blind and single blind, and the former is always recommended, especially when the primary outcome variable is apt to subjectively interfere. Only when the condition does not permit, one should adopt the single blind or open label design; if this happens, one should state the reason in the protocol.

Double blind scheme prevents all the participants in the trial from knowing the treatment assigning procedure in advance. When the primary outcome variables are evaluated subjectively (such as pain, cognitive function, obstacle grades), and the scale used is extremely apt to produce bias due to subjective factors, double blind scheme must be applied. Even if the primary outcome variable (such as biochemical index, blood pressure) is objective, in order to decrease the selection bias or subjective tendency in filling the case report, double blind scheme should be adopted. In a double blind clinical trial, both placebo and active control need to have testing reports from drug control authority, meanwhile all the contrast drugs should be in accordance with the experimental drugs in type, shape etc.

Sometimes the form or other aspects of experimental drugs and control drugs are different (often under an active control setup), say, experimental drug as tablet, and active drug as capsule; or the forms are the same but the dosages of administration are different (say, b.i.d. and q.i.d. respectively). In order to fulfill double blind scheme, double dummy method should be used. Under this circumstance, the sponsor needs to prepare a placebo that has the same appearance or dosage of administration with the experimental drug, called the placebo of experimental drug, and another placebo that has the same appearance or dosage of administration with the control drug, called the placebo of control drug. Accordingly, patients in treatment group accept both the experimental drug and the placebo of control drug; those in control group accept both the control drug and the placebo of experimental drug.

Therefore, the forms of the drugs, the times and slices of taking amount every day are the same for all the patients in the trial. This guarantees the implementation of double blind scheme.

To meet medical ethics consideration, in double blind trail, emergency letters corresponding to each patient should be prepared, and its content is the group number which the patient is assigned at random. Both the sealed envelopes and the corresponding drugs are sent to the directors of each clinical center. Only under the emergency circumstances the director of the center is allowed to open the emergency letter (such as when the serious inverse effect occurs, or the patient needs to rescue, physicians must know which kind of treatment the patient has received). Once the letter is opened and read, this case is deemed as drop out and removed from curative effect analysis, but should be included in safety analysis if the inverse effects occur. All the emergency letters should be called back with the case report at the end of the trial. The double blind state must be kept throughout the whole trial, including making the scheme, producing the random number and blind code, assigning the drugs, recording the case reports, checking the case reports by monitor, managing the data and performing the statistical analysis. Only after statistical analysis can the blind code be revealed. Any reason to reveal the blind code before the analysis is called breaking the blindness.

The double blind scheme needs strictly standard operational procedure and any unnecessary impartation of blind code is prohibited. If the proportion of breaking the blindness is too large (say, exceeds 20%), the trial will be regarded as invalid.

In Example 11.1, the double blind method is adopted so that the patient does not know whether he or she has accepted the extra Captopril treatment, even the physicians and the statisticians do not know the truth either. In order to guarantee the double blind state, the double dummy simulation technique is implemented, namely the patients in control group take the placebo that holds the same appearance as Captopril.

11.2.4 Data management

The rudimental request of a clinical trial is to guarantee that the original data and the documents are factual, scientific, normative and intact. The purpose of data management is to put the patients' data into report and database fast,

integrally and inerrably. The data management includes a series of standard operation on the case report form (CRF) that is devised according to the protocol. The physician is the first executor to fill the data; the monitor verifies the data; the data manager checks and confirms the data from CRF, and inputs the data into computer integrally and inerrably; the statistician checks the logic rationality of the data, locks the data, performs statistical analysis and writes the statistical analysis report.

11.2.4.1 *Database*

The data manager should prepare the database structure before getting the first CRF. The database needs very strong privacy and reliability. After the first CRF arrives, data manager should try to run and check the database, and make further modification if necessary.

11.2.4.2 *Further check of CRF*

The data is checked on the date of study, selection criterion, exclusion criterion, drop out, missing value etc. If there is any suspicious, a data query form should be passed through monitor to investigator; the investigator should fill in the data query form, and pass it back through monitor to data manager. The data query form should be kept properly as a sort of document in clinical trial. The content of data query form includes date of trial, title of the trial, name of center, case identification number, patient's name, etc. The main content is the question from data manager or monitor and the investigator's answer. The person who fills the data query form must sign, so as to keep evidences that can prove the revising of CRF and database, and avoid revising data artificially and arbitrarily.

11.2.4.3 *Double input*

Data from every CRF must be input into a computer database by two different persons independently; then compare the two independent data files with software package; if they are not consistent, then refer back to the primitive CRF, find out the reason and make modification. Double input can improve and enhance the consistency between the data from database and that from CRF.

11.2.4.4 *Visual checking*

When necessary, after the double input, we can print the variables in database and make a visual verification with the data in the CRF. This can further improve the complete consistency between the data of database and that of CRF.

11.2.4.5 *Computer checking*

The computer checking is implemented by a software routine written by the data manager or statistician. This check focuses on inclusion and exclusion criteria, visiting date, drop out, protocol violator, bad accident and inverse effect etc.

To make further quality control for the database, one can select at random some 5% records from all cases (when the number of all cases is less than 100, at least five cases should be selected), and then perform the visual check with CRF. If there are more than 15 errors in 10,000 data, visual checking should be implemented to all data in the database.

11.2.5 *Duty and task of statistician*

11.2.5.1 *Duty of statistician*

The statistician must be familiar with the relevant regulations and the standard operational procedures in the clinical trial. They should cooperate closely with the investigators to finish the design and statistical analysis task, guarantee the implementation of the relevant statistical requirements and guidelines in the clinical trial. The statistician should take part in the whole course of the clinical trial from the beginning to the end. The main tasks are:

- (1) Help the investigator perfecting and revising the research protocol, designing the CRF, deciding the statistical design method.
- (2) Perform randomization, double blind design and data management etc., according to the standard operation procedures.
- (3) Prepare the statistical analysis plan, finish the statistical analysis of all data, and write the statistical analysis report. Help the investigator perfecting the final report of the clinical trial.

All statistics related works that are involved in the clinical trial should be presided by qualified biostatisticians. The so-called qualified

biostatisticians refer to the professional personnel majored in medical statistics who have accepted special training and had abundant experience. They also have the ability to cooperate the investigators of the clinical trial, and to implement guidelines of clinical trial.

11.2.5.2 *Prepare statistical analysis plan*

The statistical analysis plan should be worked out by statistician and investigator together. It is more detailed than that prescribed in research protocol.

In the statistical analysis plan, the following material should be on the list: the choice of dataset used in statistical analysis, primary outcome variable, secondary outcome variables, possible data transformation method, statistical analysis method and statistical model, appraisal method for curative effect and safety, and expected statistical analysis results in the form of statistical table, etc.

The statistical analysis plan is worked out according to the protocol and CRF. Its first draft should be formed after the protocol and CRF completed. In the process of clinical trial, it can be revised, renewed and made perfect. These can be done again while blind review is going. But one must confirm it in the documentary form before the first revealment of the blind code and cannot change anymore.

11.2.5.3 *Statistical analysis of the data*

Statistical analysis methods depend on the research purpose, the research protocol and the nature of observed data. According to the statistics principle, one decides to adopt the parametric analysis or the non-parametric analysis. According to the request of statistical analysis plan, statistician compiles software routines.

Theoretically, we should do statistical analysis to all the cases randomized into the group, called intention to treat (ITT). To the incomplete observed cases, who should have experienced the whole treatment process, we carry forward the last observed data to the final result, and then do the statistical analysis.

All the data come from the patients with good compliance with the protocol compose the qualified dataset, called Per-Protocol (PP) set. The protocol violators, as the result of bad compliancy, censoring or using prohibited

11.3.2 *Random sampling, Random grouping and random ordering*

11.3.2.1 *Random sampling*

“Random sampling” means taking a certain amount of individuals from a finite population to a sample.

Example 11.2 There are 4600 students in a school. If we want to estimate the myopia rate, we will investigate a sample of 5% (230) students.

Firstly, we get a numbered list for the whole school students from 1 to 4600; secondly, we generate 230 random numbers between 0 and 1 by computer, multiply 4600 and take integers, abandon the repeated number if any and re-sample; finally, the investigator may select the corresponding students in the numbered list and constitute the sample.

11.3.2.2 *Random grouping*

Example 11.3 20 animals are included in an experiment. How to assign them at random to group A or B?

Firstly, number animals from 1 to 20 according to their weights; secondly, set the seed numbers as 25683 and generate 20 different random numbers by SAS (release 6.12); the first ten numbers correspond to group A, else to group B; then, map the animals' serial number to the rank of the random numbers above (see Table 11.1); finally, arranging the animals' serial number with the corresponding groups, get the allocation decision (see Table 11.2).

11.3.2.3 *Random ordering*

Example 11.4 There are five animals in an experiment, apply the random ordering method to determine the order for them to accept the treatments.

Firstly, number the experiment animals: 1, 2, 3, 4 and 5. Set the seed number as 8888 and generate five random numbers in the SAS (release 6.12): 0.94732, 0.1485, 0.63843, 0.53516 and 0.20371; then match 1, 2, 3, 4 and 5 with the rank of 0.94732, 0.1485, 0.63843, 0.53516 and 0.20371, the order for the animals 1, 2, 3, 4 and 5 to accept the treatments is 2, 5, 4, 3 and 1 respectively.

Table 11.1 20 animals' random allocation.

Experiment objects' numbers	Group	Arranged by ranks of random numbers
1	A	
2	B	
3	B	
4	A	
5	B	
6	B	
7	B	
8	B	
9	A	
10	B	
11	A	
12	A	
13	B	
14	A	
15	A	
16	B	
17	A	
18	A	
19	A	
20	B	

Table 11.2 20 animals' random numbers.

Group	Random numbers	Ranks of random numbers
A	0.79989	19
A	0.79228	17
A	0.48209	9
A	0.63182	15
A	0.79402	18
A	0.00467	1
A	0.61974	14
A	0.51783	12
A	0.11899	4
A	0.51161	11
B	0.16158	5
B	0.49511	10
B	0.19743	6
B	0.10744	3
B	0.96165	20
B	0.54952	13
B	0.44437	8
B	0.68038	16
B	0.06685	2
B	0.40640	7

11.3.2.4 Stratified randomization

Generally speaking, the randomization allocation can enhance the balance among the contrast groups but does not necessarily guarantee the distributions of important confounding factors balanced among the contrast groups. Sometimes we should stratify the important confounding factors that may have important impacts on the experiment results, and then carry on randomization in each stratum. This method is called stratified randomization. For example, in order to guarantee the same gender proportion between treatment and control groups, we should carry on randomization for male and female respectively. Another example in multi-center clinical trials is that after the patients being stratified by center, the randomization allocation is performed within each center. One more example, sometimes the entire duration of a clinical trial, from the beginning till the entering of all the patients, needs more than half a year. Under this situation, accounting for

Table 11.3 Allocation scheme of block randomization.

Block	Treatment				
	A	B	C	D	E
1	5	1	3	2	4
2	4	1	5	2	3
3	4	3	5	1	2
4	3	2	5	4	1
5	2	4	3	5	1
6	3	2	1	5	4

possible impact of seasonal factor, we should apply stratified randomization in each time interval to improve the balance between contrast groups in terms of entering time.

In block design, the experiment subjects are randomized within blocks. For example, if there are 30 experiment subjects, who have been equally divided into six blocks, according to an important confounder, then one can randomly assign the subjects in each block into five treatment groups by the above-mentioned random ordering method (see Table 11.3).

In order to ensure the reliability of experiment, one should document randomization method, random numbers, seed number and the program routine for generation of random numbers. If one gets the random numbers from a random number table, he or she should describe the source of the random number table used, the starting row, column and page etc. In the clinical trial for new drugs, the random numbers should be replicable. The parameters and program routines for generating random numbers should be sealed up with blind code together. If the random numbers are generated by computer, one should explain and declare the software, program and seed number etc.

11.4 Randomized Controlled Trial

11.4.1 Definitions and characteristics

Randomized controlled trial (RCT) is a type of clinical trials using prospective method to determine the treatment effect by comparing the outcome of intervention group and control group. RCT is applied not only in the field of

clinical research, but also in behavioral intervention, evaluation of screening programs and so on. In RCT, various known or unknown confounders can be balanced by randomization, so that the average difference in baseline can be lower or even eliminated.

In addition to randomization, RCTs have more important characteristics:

- (1) **Intervention:** RCT is often performed for a comparison through intervention. The intervention group in a RCT could receive a predesigned measure of therapy or prevention, while the control group only receives a placebo; or the intervention group receives a most concerned intervention while the control group receives another intervention or some standard measure.
- (2) **Prospective:** RCT is a special kind of prospective researches. It does not require each subject to be followed up starting from the same point as long as the starting and ending points of the follow-up being clearly defined.
- (3) **Data analysis:** After randomization, if the baseline values of important variables other than the treatment between the compared groups are overall balanced, uni-variate statistical methods can be used for data analysis. Otherwise, if the baseline values between the two groups are somewhat unbalanced, stratified analysis or adjusted analysis should be adopted, which should be specified in the statistical analysis plan in advance. A principle of intention-to-treat (ITT) should be followed in data analysis for efficacy that all subjects' data should be analyzed in accordance with the initial grouping by randomization, regardless of whether the subjects accept any other interventions. If the analysis for efficacy is conducted only based on the data of those subjects, who actually follow the allocation by randomization, which is called per-protocol analysis, one has to clarify and explain in the report.
- (4) **Ethical issues:** Similar to other researches involving human subjects, signed informed consent should be obtained before the trial starts. Subjects may be randomly allocated to intervention group or control group. The treatments for compared groups should meet the principles of uncertainty and equipoise that the real pros and cons of each treatment cannot be determined by the individual researcher and the medical organizations. Otherwise, the RCT will offend the ethical standards.

11.4.2 *Classification*

RCTs can be divided into different categories based on different focus of classification. As mentioned in Sec. 11.2, RCTs can be divided into four phases of clinical trials according to the stage; RCTs can be divided into N of 1 trial, sequential trial and fixed trial according to the design; and according to the purpose, RCTs can also be divided into explanatory trial and effectiveness trial.

N of 1 trial is a special kind of RCTs. By using single case, multi-round crossover trial is repeatedly tested on the same individual to evaluate the difference of individual effects caused by it.

For a sequence trial, there is no need to determine the sample size in advance. The subjects enter the study in sequence, and the statistical analysis is continuously conducted with the size increasing until the conclusion is obtained. Opposite to this, the trials which have sample size decided during design stage are known as fixed trials.

The explanatory efficacy trial, also known as efficacy trial, aims at understand whether the treatment/intervention is effective. It tries to remove all bias and confounding factors though the treatment effect obtained in such ideal condition may not reflect the real effectiveness in clinical practice.

Pragmatic trial is also known as effectiveness test. It tries to investigate if the treatment/intervention is effective in clinical practice and to estimate its actual effect size. The implementing condition of the trial is identical to that of the real clinical practice; hence, the finding on the treatment effect can be referenced for clinical decision-making in the practical settings.

The detailed difference between explanatory trial and pragmatic trial can be seen in Table 11.4.

11.4.3 *Quality assessment*

Since RCT can provide the strongest evidence for causal association comparing with other trials, and the results of RCT often influence the health decision-making, both quality and quantity assessment of RCT are very important. Quality of RCT includes methodological quality and reporting quality. Methodological quality explains whether the study design, implementation and analysis can avoid or reduce the occurrence of bias as much as possible. It also reflects how the design and implementation match to

Table 11.4 The difference between explanatory trial and pragmatic trial.

	Explanatory trial	Pragmatic trial
Objectives	To evaluate efficacy	To compare effectiveness
Filed sites	The environment similar to laboratory; to control the background factors as much as possible	Actual environment; similar to the real situation in clinical or community setting as much as possible
Subjects	Strict inclusion and exclusion criteria; high homogeneity and low catholicity; allow small sample size	Include from the actual target population with weak exclusion criteria; low homogeneity and high catholicity; large sample size
Intervention	A particular treatment which should be implemented strictly	Conventional treatment which can be adjusted according to subjects' status
Control	Placebo as control in order to explore the efficacy of the intervention	Conventional treatment as control to find the optimal strategy
Blinding	Usually use double blinded method	No requirements on blinding, but the confidentiality of random allocation should be ensured. If subjectivity exists in data collection, people who report or collect data should be kept blind of allocation.
analytical method	Intention to treat analysis	Pre-protocol analysis
outcomes	Sole laboratory measurement with high specificity	Multiple outcomes which reflect the actual status of subjects
Follow-up period	Short	Long

the study objectives. Reporting quality focuses on whether the design, implementation and analysis are reported appropriately. The inadequate research report will affect the interpretation and application of the results. The commonly used assessment tools include Jadad scores, Delphi list and Cochrane bias scale *et al.* These assessment tools give a score to each question, and determine the level of test quality by the total score.

Jadad score is a tool used independently to evaluate the methodological quality of clinical trial. It judges the quality of the research design by grading the situation of randomization, blind and withdraw/lost (J Clin Epidemiol.

1998;51:1235-41; <http://www.cochrane-handbook.org/>). Similar to Jadad score, Delphi list and Cochrane bias scale adopt scoring method to judge the quality of the experimental design (J Clin Epidemiol. 1998;51:1235-41; <http://www.cochrane-handbook.org/>).

11.4.4 Report

In order to standardize the reports of RCTs, to increase transparency in reporting methods and results, and to guide the design of RCTs, the CONSORT (Consolidated Standards of Reporting Trials) working group (which consists of experts in clinical trials, statisticians, clinical epidemiologists, and biomedical experts and so on) proposed a “CONSORT STATEMENT” in 2001, and updated in 2010 with the accumulation of new methodological evidence and experience.

The content of CONSORT 2010 STATEMENT included a checklist and a flowchart. The checklist contains 25 items and made requirements on title, abstract, methods, results, discussion and sponsorship, etc. The flowchart requires the researchers to report sample size in every step of the trial (including subject recruitment, randomization, follow-up, and data analysis) and to explain the reason of changes in sample size. CONSORT 2010 STATEMENT, although focuses on the parallel randomized controlled trials, can also be referenced by other types of randomized controlled trials. More information of the STATEMENT can be obtained from the CONSORT website: <http://www.consort-statement.org>.

11.4.5 Registration

In 2004, the ICMJE (International Committee of Medical Journal Editors) declared that since July 1, 2005, any RCT, of which the findings are intended to be published in the journals of the ICMJ should register before the first case is recruited. This requirement aims to reduce publication bias caused by less publication of negative results.

RCTs can be registered through the website established by the U.S. National Institutes of Health (NIH) via National Library of Medicine (NLM) (<http://clinicaltrials.gov>), the website (<http://www.who.int/ictrp/en/>) established by International Clinical Trials Registry Platform (ICTRP) or

the website (<http://isrctn.org>) established by International Standard Randomized Controlled Trial Number (ISRCTN).

11.4.6 *The subject's preferences*

According to the principle of informed consent, the subjects of RCTs should be informed about the procedures on allocation, treatment and others before they join the study. However, in practice, the participants may have their preference on certain treatment. Some qualified subjects may refuse to receive conventional or experimental treatment when they know that they will be allocated randomly. This selective refusing caused by treatment preference would affect the comparability of the treatments.

If these subjects with preference are randomly allocated without blinding, larger bias would be generated. If the subjects are allocated to the group which they do not want to join, the subjects may have poor compliance because of inner discontent. Such negative attitude may lead to worse treatment effect when comparing with those without preference. In contrast, if the subjects are allocated to their preferring group, their compliance and the treatment effect may be better than those without preference. Considering the problem caused by preference, researchers proposed several different study designs.

(1) **Comprehensive cohort design** When subjects who accept randomization accounted for a relatively small proportion of all eligible subjects (e.g. <30%), all subjects should undergo follow-up. Whether or not accept randomization can be treated as one of the covariates in data analysis. Hence, the external validity and the extent to which the results can be generalized can be assessed. However, it should be noted that, such design cannot take the place of RCT design; moreover, the sample size should be large enough to ensure the statistical power.

(2) **Patient-preference design** Some researchers suggested that the preferences of subjects should be known before randomization, and subjects with preferences can be allocated to their preferred group, while subjects without preference can be allocated to the group according to randomization. When comparing the effects of treatment *A* and *B*, four parallel groups (preferred group *A*, randomized group *A*, preferred group *B*, randomized group *B*, Fig. 11.2) should be followed up, respectively.

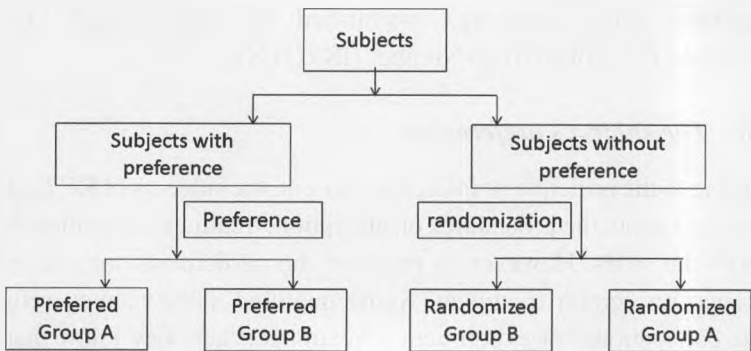


Fig. 11.2 Grouping diagram of patient preference design.

However for such kind of design, there is no sophisticated statistical analysis method available currently. The data from randomization groups can be analyzed directly, but it is not suitable to analyze the data of those preferred groups directly. For example, in a comparative study of termination of pregnancy surgery, the pregnant women who met the inclusion criteria can select the ways of termination of pregnancy (mifepristone drugs or vacuum aspiration) according to their preference. In 363 pregnant women, about half (168) selected different ways according to their preference; the remaining 195 women who had no preference were randomly allocated into two groups. A follow-up survey was conducted to study the acceptability of the two methods. Researchers found that for group with preference, there is no statistically significant difference in the acceptability between medication method and vacuum aspiration. However, for the group without preference (randomized group), the results was statistically different. Only two women in the vacuum aspiration group would choose the other method in future but 21 in the medication group wanted to select the other method in future. With the help of patient preference design, the information which could not be discovered in conventional RCT was reported.

11.5 Comments on Some Medical Examples

Example 11.5 An article entitled *Primary observation of 735 cases of cesarean section combined with implantation of intrauterine device synchronously* (from Chinese Journal of Gynecology and Obstetrics, 1985;

20(1): 49–50). In the article, the author collected 1562 cases undergone cesarean section. Among them, 735 cases acted as experimental group, which were implanted intrauterine device synchronously with cesarean section, and all of the 735 cases met the following indications: rupture of membranes and total stage of labor were no more than 24 hours, no infection (no positive signs were found and peripheral blood routine examination showed normal), and acquired the agreement of the parturient. The other 827 cases were treated as control group without implantation of intrauterine device during the labor procedure. The bleeding condition, the time of lochia stopping, the adverse effects and other variables after operation were compared.

In this study, not only the treatments on the two groups were different but the baselines of the two groups were evidently unbalanced, the experimental group was better, and the control was worse. These two groups were incomparable.

This kind of mistakes is common in clinical trials. The problem is that the authors divide the well-conditioned consenter into experimental group, and gather the patients to the control group who are unwilling to accept the new therapy, or have no indications for the treatment, or cannot afford the cost, so as to make the two groups own the different baseline, and become incomparable. In this example, the author should randomize the 735 cases of female who had the stated indications into two groups, and implant the intrauterine device to the cases of one group and do not to those of another group.

Example 11.6 In another article entitled *Study on the risk factors of breast cancer* (from Chinese Journal of Epidemiology, 1981, 2(4): 253), the author recruited 607 pairs of case and control according to the requirement that the difference of ages within the pair was under five years. And then, without hypotheses testing, the author thought that the age distribution of the two groups were similar. But by comparing the age distributions of the two groups (Table 11.5), we conclude that they are significantly different ($\chi^2 = 17.25$, $P = 0.004$).

This kind of mistake is common in the field study and laboratory study. Although the matching requirements were considered and strictly prescribed, the case and control groups were still incomparable because of the systematic error in the executing procedure, or because the requirements

Table 11.5 Age distribution of 607 pairs of subjects observed.

Age	No. of cases	No. of control
20	3	6
30	72	84
40	193	244
50	228	199
60	101	67
70	10	7

were not strictly met, and led to the unbalance in the two groups. Now the difference of the age distributions in two groups has already been significant. To solve the problem, method of adjusting age difference is needed in the analysis, such as the Mantel Haenszel stratification analysis or multiple logistic regression.

Example 11.7 A clinical trial was to study the effect of injection A on inflammation with injection B as control. The researcher chose genitourinary infection and dental inflammation as indications. Three groups were planned, one is experimental group, another is control group, and the third is open group. The researcher collected 60 cases for the experimental and control group respectively, among which there were 30 cases of genitourinary infection and 30 cases of dental inflammation; but the open group contained 28 cases of surgical infections and 15 cases of trichomonas vaginitis. To increase the sample size for the treatment of injection A, the researcher combined the experimental group and the open group, and compared them with the control group.

Between the experimental group and control group, the case number and disease type were well matched, but the number of cases in the experimental group was insufficient. In the open group, although the injection was the same as experimental group, the disease type was different from both the experimental and control groups. If there are no cases of the same kinds of diseases as in the open group, the control group failed to play the role of control. In the same way, it could not be compared to the combination group.

This kind of design is a common mistake in the clinical trials for new drug. On the one hand, a large sample size is required; on the other hand, in order to improve the marketing of the new drug, more indications should be

included and so the open group is applied. In fact, the method of combination is rarely used except for extremely special variables such as adverse effect rate.

In all the three examples above, the assigning of control group was problematic, and the key point is the balance among the groups. To correct the mistakes, in Example 10.7 the open group should be eliminated; in Example 10.6 age should be regarded as a confounder and its effect needs to be adjusted by statistical technique; however, Example 10.5 was unrepairable.

11.6 Computerized Experiments

Experiment 11.1 Generating ten random numbers between 0–999 The software SAS supplies the function RANUNI() to generate random numbers, by putting a seed in the bracket. Different seed produces different series as the following program shows.

In program 11.1, lines 01 to 10 generate data AAA; lines 02 and 03 appoint two different random seeds, lines 05 and 06 generate different random number series based on the above seeds respectively, lines 07 and 08 take the integer of (random number \times 1000), line 09 delivers the output to data AAA, lines 04 and 10 repeat ten times of lines 05–09 to generate two series of ten random numbers. Lines 11 and 12 print the data AAA.

Experiment 11.2 Randomized Grouping Assigning 20 animals into groups *A* and *B* randomly.

In program 11.2, lines 01–05 generate 20 random numbers, *X* with id numbers 1 to 20, in data AAA. Lines 06–08 generate SAS dataset BBB,

Program 11.1 Generating random number.

Line	Program	Line	Program
01	DATA AAA;	07	Y1=INT(X1*1000);
02	SEED1=20000720;	08	Y2=INT(X2*1000);
03	SEED2=20041012;	09	OUTPUT;
04	DO I=1 TO 10;	10	END;
05	X1=RANUNI(SEED1);	11	PROC PRINT;
06	X2=RANUNI(SEED2);	12	RUN;

Program 11.2 Completely randomized grouping.

Line	Program	Line	Program
01	DATA AAA;	10	SET BBB;
02	DO ID=1 TO 20;	11	UNIT=RX;
03	X=RANUNI(26853);	12	IF ID<=10 THEN TRTMEN='A';
04	OUTPUT;	13	ELSE TRTMEN='B';
05	END;	14	KEEP UNIT TRTMEN;
06	PROC RANK OUT=BBB;	15	PROC SORT;
07	RANKS RX;	16	BY UNIT;
08	VAR X;	17	PROC PRINT NOOBS;
09	DATA CCC;	18	RUN;

in which all the variables in data AAA and the ranks of random variable X, namely variable RX, are included. Lines 09–14 generates data CCC and assigns the 20 experiment subjects randomly into two treatments: line 10 copies all the variables in data BBB into data CCC; line 11 makes RX equal to UNIT; lines 12 and 13 assign 20 random numbers into two treatments according to their IDs; line 14 drops out other variables except for UNIT and TRTMEN. Lines 15 and 16 sorts data CCC. Lines 17 and 18 print data CCC sorted by UNIT.

Experiment 11.3 Randomized block grouping Assign the 30 subjects into six blocks according to their similarity, each block includes five subjects.

In program 11.3, lines 01–07 generate 30 random numbers into six blocks with five numbers in each block in data AAA. Lines 08 and 09

Program 11.3 Randomized block group design.

Line	Program	Line	Program
01	DATA AAA;	10	PROC RANK OUT=BBB;
02	DO BLOCK=1 TO 6;	11	RANKS TRTMEN;
03	DO UNIT=1 TO 5;	12	VAR X;
04	X=RANUNI(37277);	13	BY BLOCK;
05	OUTPUT;	14	PROC SORT;
06	END;	15	BY BLOCK TRTMEN;
07	END;	16	PROC PRINT NOOBS;
08	PROC SORT;	17	RUN;
09	BY BLOCK ;		

Program 11.4 Plan for stratified randomization design.

Line	Program	Line	Program
01	DATA AAA;	13	PROC RANK OUT=BBB;
02	CENTER=3;	14	RANKS RX;
03	B=4;	15	VAR X;
04	T=2;	16	BY STRATA BLOCK;
05	DO STRATA=1 to CENTER;	17	DATA CCC;
06	DO BLOCK=1 to B;	18	SET BBB;
07	DO TRTMENT=1 to T;	19	NUMBER= _n_;
08	X=RANUNI(37277);	20	IF RX=1 THEN GROUP='A';
09	OUTPUT;	21	IF RX=2 THEN GROUP='B';
10	END;	22	KEEP STRATA BLOCK NUMBER GROUP;
11	END;	23	PROC PRINT NOOBS;
12	END;	24	RUN;

sort data AAA by block. Lines 10–12 generate data BBB, in which the ranks of every five random numbers in each block are stored into variable TRTMENT. Lines 14 and 15 sort data BBB according to BLOCK and TRTMENT. Line 16 prints the result with every unit corresponding to different treatment within each block.

Experiment 11.4 Stratified Randomized Grouping Divide 24 subjects into three strata, each stratum comprise four blocks, each block has two paired subjects, to whom the two treatments are randomly assigned respectively.

In program 11.4, lines 01–12 generate 24 random numbers with three strata, each stratum comprising four blocks, each block having two paired subjects, into data AAA. Lines 13–16 generate data BBB, in which the ranks of every two random numbers in each block and stratum are stored into variable RX. Lines 17–21 assigns every two subjects in each stratum and block into group A and group B according to their ranks. Line 22 drops out temporary variables and lines 23 and 24 prints the result.

11.7 Practice and Experiments

1. How many types of experiment errors are there, what are they? How to control the experiment errors during the stage of experiment design?

2. What are the purposes of assigning control group, replication and randomization respectively?
3. What are the advantage and disadvantage of randomized block design? What are the differences between completely randomized design and randomized block design? In what kind of situations are they suited?
4. An institute organized a clinical trial on a new drug. There were in total 200 patients in this trial, now to equally divide them into two groups for the new drug and control respectively, how many possible designs can we choose? Perform the corresponding designs by computer.
5. Suppose there are 24 male rats with similar weights, try to divide them into three groups by completely randomized design.
6. Suppose there are six nests with five animals each. Try to divide them into five groups by randomized block design.
7. 100 patients took part in a clinical trail, according to their order of hospitalization, the first 50 patients were assigned into group *A*, and the following 50 patients were assigned into group *B*. What do you think of this design, does it violate the statistical principles? Give your reason.
8. (Continued of 7) If the first hospitalized patient was assigned according to certain randomization schedule, say, into group *A*, then the second patient must goes to group *B*, the third patient to Group *A*, fourth patient to Group *B* . . . and so on. Do you think that this design is in agreement with relevant statistical principles? Why?
9. In order to study the effect of Lysine (Lys) on children's growth, the researchers plan to add Lysine in nursery children's bread as intervention.
 - (1) In this study, how to set up a control group?
 - (2) What are the observational indices?
 - (3) Which non-treatment factors should be controlled? How?

(1st edn. Tong Wang, Qinghai Liu, Zhen Yang, Jiqian Fang; 2nd edn. Jing Gu, Jiqian Fang)

Part II

Multi-variate Statistics



Chapter 12

Multiple Regression and Correlation

In medical research, like in all other areas, the response variable may depend on more than one variable. For example, a person's blood pressure depends not only on the person's age, but also on the person's gender, workload, eating and drinking habits, smoking and family history, etc. The extension of the simple regression, where there is only one independent variable in the regression equation, to a regression equation that delineates the dependence of the mean of a random dependent variable Y on several independent variables, X_1, X_2, \dots, X_p , is called "multiple linear regression", or simply "multiple regression". If the values of the independent variables are not by design, but concomitant with Y , we use multiple regression to find the conditional mean of Y given X_1, \dots, X_p , which is a linear function of X_1, \dots, X_p , or apply multivariate methods to the $(p + 1)$ variables (Y, X_1, \dots, X_p) , such as multiple correlation and partial correlation.

12.1 Basic Procedure of Multiple Regression

12.1.1 *The model*

Given p independent variables, X_1, X_2, \dots, X_p and the dependent variable Y , the extension of the simple regression to the present situation is the following multiple regression equation:

$$\mu_{Y|X_1, X_2, \dots, X_p} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p, \quad (12.1a)$$

where α has the same meaning as that in the simple regression and is called the constant term, or the intercept, of the regression equation. $\beta_1, \beta_2, \dots, \beta_p$ are the partial regression coefficients or simply regression coefficients. β_i is the increment in the mean of Y per unit increase in X_i when other

independent variables are fixed. We further assume that the conditional variance of Y given X_1, \dots, X_p is a constant, not depending on the values of the X variables:

$$\sigma_{Y|X_1, \dots, X_p}^2 = \sigma^2. \quad (12.1b)$$

For a sample of n individuals with measurements $(x_{i1}, x_{i2}, \dots, x_{ip}, y_i)$, $i = 1, 2, \dots, n$, we have the fitted equation

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \dots + b_p X_p, \quad (12.2a)$$

where b_0, b_1, \dots, b_p are the estimates of $\beta_0, \beta_1, \dots, \beta_p$. These estimates are obtained from the least-square method, such that the choice of b_0, b_1, \dots, b_p will make the sum of squares of differences between the observed y_i and the fitted \hat{y}_i attaining the minimum. Unlike the simple regression, the computation for multiple regression is more tedious and nowadays we can rely on computing software.

Very often, the variables are not commensurable, because of different levels of measurement scale or different nature of measurements. In such cases, the regression coefficients are not directly comparable, hindering comparative interpretation of the regression coefficients. One way out is to standardize all the variables, i.e. let $Y' = (Y - \bar{Y})/S_Y$ and $X'_i = (X_i - \bar{X}_i)/S_i$, and fit the regression equation. The estimates thus obtained are denoted by a', b'_1, b'_2, \dots , called the standardized regression coefficients. As the means of standardized variables, \bar{Y}' and $\bar{X}'_1, \bar{X}'_2, \dots, \bar{X}'_p$ are all zero, we have

$$a' = \bar{Y}' - b'_1 \bar{X}'_1 - \dots - b'_p \bar{X}'_p = 0.$$

The corresponding equation is

$$\hat{Y}' = b'_1 X'_1 + b'_2 X'_2 + \dots + b'_p X'_p. \quad (12.2b)$$

The relationship between b_i and b'_i is

$$b_i = b'_i S_Y / S_i, \quad b'_i = b_i S_i / S_Y.$$

The standardized regression coefficients can be directly compared to show which independent variable has the greatest effect, etc.

Table 12.1 Data from the second hospital of Chong Qing medical university 1970–1989.

Year	In-patients Y	Out-patients (10 K) X_1	Bed-usage (%) X_2	Turnaround X_3
1970	6349	49.8	94.25	19.84
1971	6519	38.1	98.50	20.37
1972	5952	36.6	89.86	18.80
1973	5230	36.0	86.00	16.34
1974	5411	32.3	83.29	16.91
1975	5277	37.8	77.88	18.07
1976	3772	34.1	92.62	17.96
1977	3846	42.2	86.57	18.31
1978	3866	38.1	84.29	18.41
1979	5142	39.5	89.29	20.61
1980	7724	55.8	97.63	21.72
1981	8167	63.0	96.53	23.33
1982	8107	65.2	93.43	21.91
1983	7998	66.1	94.45	21.05
1984	7331	65.4	93.03	19.96
1985	6447	60.1	91.79	18.81
1986	4869	56.9	88.94	15.82
1987	5506	57.7	91.79	16.01
1988	5741	53.4	99.03	16.59
1989	5568	48.7	94.93	19.09
Average	5941	48.8	91.21	18.99

12.1.2 Basic procedure with an example

Example 12.1 Shi Lei (1991) published the annual number of out-patients X_1 , percentage of bed usage X_2 , bed turnaround X_3 and number of in-patients Y in Table 12.1 in his hospital during 1970–1989, and fit a regression equation for Y in terms of X_1 , X_2 , X_3 . Here we use the data to illustrate the multiple regression analysis using SAS.

Let us first look at the analysis of variance table for the regression model, generated by the SAS output as given in Table 12.2. One can see from the table that the partition of degrees of freedom and the sum of squares is similar to the simple linear regression where $p = 1$. The Corrected Total Sum of Squares (SS_{Total}) is just the sample sum of squares corrected for the mean for the values of Y and hence $SS_{\text{Total}}/(n - 1)$ is an unbiased estimate of the

Table 12.2 Analysis of variance for regression.

Source	<i>DF</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Pr > F</i>
Model	3	27066405	9022135	15.26	<0.0001
Error	16	9461837	591365		
Corrected Total	19	36528242			

marginal (unconditional) variance of Y . The conditional mean of Y given by (12.1a) is a random variable as a function of the X variables and thus has a marginal variance, denoted by $Var(\mu_{Y|X})$, which will become zero if all the coefficients of the X variables in (12.1a) are zero. Now the marginal (unconditional) variance of Y is a sum of two parts: $Var(\mu_{Y|X})$ and σ^2 , the conditional variance in (12.1b). Thus the expected value of SS_{Total} is $(n - 1)(Var(\mu_{Y|X}) + \sigma^2)$. The expected value of the Sum of Squares of Error (SS_{Error}) is $(n - 1 - p)\sigma^2$, and hence the expected value of the Regression Model Sum of Squares ($SS_{\text{Regression}}$) is $(n - 1)Var(\mu_{Y|X}) + p\sigma^2$. The SS_{Error} is the same as the least-square that has been achieved when minimizing the sum of squares of deviations between the observed and the fitted values of Y in order to obtain the least-square estimates of the regression coefficients, hence also called residual sum of squares denoted by $SS_{\text{Residual}} = SS_{\text{Error}}$. In the table, all the Mean Squares equal the Sum of Squares divided by the corresponding DF (degree of freedom). Thus the Mean Square for Error (MS_{Error}) estimates σ^2 and the Mean Square for the regression model ($MS_{\text{Regression}}$) estimates $\sigma^2 + (n - 1)Var(\mu_{Y|X})/p$. If the conditional mean of Y does not depend on the X variables, or equivalently if the following null hypothesis is true,

$$H_0: \beta_1 = \beta_2 \cdots = \beta_p = 0,$$

then $MS_{\text{Regression}}$ and MS_{Error} will have the same expected value and thus the ratio $F = MS_{\text{Regression}}/MS_{\text{Error}}$ will be very close to one. As a random variable, this ratio has an F distribution with $v_1 = p$ and $v_2 = (n - 1 - p)$ degrees of freedom under the null hypothesis. If the observed F statistic is larger than the tabulated F critical value, or if the p -value of the statistic is smaller than the chosen significance level, we have doubts about the F distribution and hence reject the null hypothesis. In this particular example, we have $F = 15.26$, much larger than one; indeed much larger than the

Table 12.3 Parameter estimates and tests.

Variable	DF	Parameter estimate	Standard error	<i>t</i>	Pr > <i>t</i>	Standardized estimate
Intercept	1	-4848.94404	3128.70698	-1.55	0.1407	0
X_1	1	55.88633	18.00125	3.10	0.0068	0.47781
X_2	1	21.92976	39.81302	0.55	0.5894	0.08708
X_3	1	319.04673	96.59315	3.30	0.0045	0.48403

critical value, 3.24, from the F table with $\nu_1 = 3$, $\nu_2 = 16$ ($n = 20$, $p = 3$). Equivalently, the p -value from the SAS printout is less than 0.0001, which in turn is less than any commonly used significance level ($\alpha = 0.01$, $\alpha = 0.05$, or $\alpha = 0.10$). So we reject the null hypothesis. That is, at least one of the regression coefficients is not zero, but we do not know which one.

Next we look at the estimates of regression coefficients and their standard errors given in Table 12.3. According to the table, the fitted regression equation is

$$\hat{Y} = -4848.94 + 55.89X_1 + 21.93X_2 + 319.05X_3. \quad (12.3)$$

The t statistic and its p -value for testing the corresponding regression coefficient are also given in Table 12.3. At the usual level, $\alpha = 0.05$, the effects of the number of out-patients ($p = 0.0068$) and frequency of bed turnaround ($p = 0.0045$) are statistically significant in predicting the number of in-patients, but the percentage of bed-usage ($p = 0.5894$) is not. The Standardized Estimates in the last column of the table show the direction and relative magnitude of the impact of the independent variables on Y , the larger the absolute value the more the impact. In this case, all independent variables will have positive impact on the value of Y , with bed-turnaround (0.48403) the most influential, closely followed by the number of out-patients (0.47781). The percentage of bed-usage (0.08708) has very little effect. It should be noted that the p -value is not meant to indicate the impact of an independent variable on Y like what the standardized estimate does, but only the amount of risk involved in judging that the impact is real and not merely by chance.

There are two statistics that indicate the goodness-of-fit for the equation to the data, the root mean square error (RMSE) and the multiple determination coefficient, denoted by s and R^2 respectively. The smaller the s ,

the better the fit. As a natural estimate of σ , s is also called the residual standard deviation. It always decreases when we add more X variables in the equation, here in this example $s = \sqrt{591365} = 769$. The R^2 equals the square of the simple correlation coefficient between the observed and the fitted values of Y . It is interpreted as the proportion of the variation of Y values that the model can explain, because it can be expressed as

$$R^2 = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}}.$$

Here $R^2 = 0.7410$, meaning that about 74% of the variability of Y is attributed to its dependence on the number of out-patients, hospital bed-usage percentage and turnaround frequency, pertaining to the quantity $Var(\mu_{Y|X})$ in the discussion on Table 12.2.

As in the simple linear regression, the SAS output provides the predicted value of Y , its standard error, the confidence limits and prediction limits for all sets of values of the independent variables which have been used in the model fitting as given in Table 12.4. If there is a missing value of Y corresponding to a complete set of values for the independent variables, the SAS will still provide prediction for the missing value of Y . This feature of SAS allows automatic prediction of Y for any given sets of values of the independent variables. In Table 12.4, the first line means that for $X_1 = 49.8$, $X_2 = 94.25$ and $X_3 = 19.84$, the estimated conditional mean and the predicted value of Y are both 6331, but the 95% confidence interval for the conditional mean is (5893, 6769), while the 95% prediction interval for a particular year is (4643, 8019). We can predict with 95% confidence that for $X_1 = 49.8$, $X_2 = 94.25$ and $X_3 = 19.84$, the number of in-patients is

Table 12.4 Confidence intervals and prediction intervals of Y .

No.	VarY	Predicted value	Std error of mean	Lower 95% of mean	Upper 95% of mean	Lower 95% predicted	Upper 95% predicted
1	6349	6331	206.5591	5893	6769	4643	8019
2	6519	5939	451.5272	4982	6897	4049	7830
.
19	5741	5600	467.8839	4608	6592	3692	7508
20	5568	6045	226.1359	5566	6525	4346	7744

Table 12.5 Correlation matrix for Example 12.1.

CORR	X_1	X_2	X_3	Y
X_1	1.0000	0.5317	0.4065	0.7209
X_2	0.5317	1.0000	0.4570	0.5623
X_3	0.4065	0.4570	1.0000	0.7181
Y	0.7209	0.5623	0.7181	1.0000

between 4643 and 8019. Note that the prediction interval is wider in general because the prediction of Y equals the estimated conditional mean plus a prediction from a normal distribution with zero mean and an estimated $\sigma^2 = 591365$.

12.2 Multiple Correlation

We shall consider correlation analysis only when X_1, \dots, X_p and Y jointly follow a multivariate normal distribution.

12.2.1 Simple correlation coefficient

When data are given for the variables X_1, \dots, X_p and Y , we can compute as before the simple correlation coefficient between any pair of the $(p + 1)$ variables and arrange the simple correlation coefficients in the form of a correlation matrix. The correlation matrix for Example 12.1 is given in Table 12.5.

Note that the diagonal elements in the correlation matrix are all ones and all other elements are symmetric about the diagonal. So we can either retain the upper diagonal part or the lower diagonal part of the matrix. We observe that the simple correlation between X_1 and Y , 0.7209, is the largest and the next largest is between X_3 and Y , 0.7181. Thus the simple linear regression of Y on X_3 gives $R^2 = (0.7181)^2 = 0.5157 = 52\%$ and that of Y on X_1 gives $R^2 = (0.7209)^2 = 0.5197 = 52\%$, the two answers are quite close.

12.2.2 Multiple correlation coefficient

The linear association between a random variable Y and a set of random variables (X_1, \dots, X_p) is measured by the so-called multiple correlation

coefficient. Let the fitted regression equation be

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p.$$

The multiple correlation coefficient is defined as the simple correlation coefficient between Y and \hat{Y} , which is invariably positive; that is,

$$R = \text{corr}(Y, \hat{Y}).$$

Thus the squared multiple correlation coefficient equals the multiple determination coefficient, which is defined in the previous section.

12.2.3 Partial correlation coefficient

In Problem 5 of Chap. 9, we have considered the simple correlation coefficients of the scores in language X , mathematics Y and IQ Z . Generally, students with higher IQ would tend to do better in language and mathematics. How would X and Y associate after removing the effects of Z ? In Fig. 12.1, (a) shows the simple correlation between X and Y , quite strong; (b) and (c) show the regressions of X and Y on Z and the corresponding residuals respectively; (d) shows the simple correlation between the two sets of residuals, very little correlation indeed. The simple correlation coefficient between the two sets of residuals is called the partial correlation coefficient between X and Y after removing (or adjusting or controlling) the effect of Z , denoted by $r_{XY \cdot Z}$.

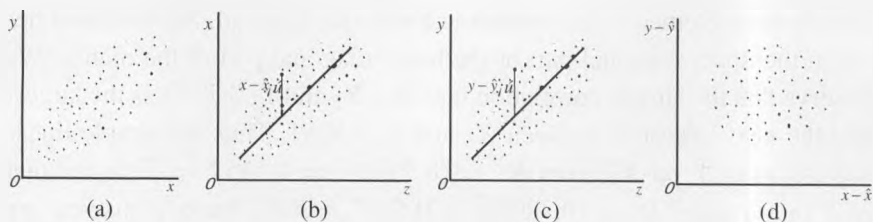


Fig. 12.1 Partial correlation between language score (X) and mathematics score (Y) after removing the effect of IQ (Z): (a) simple correlation between X and Y ; (b) regression of X on Z and its residual; (c) regression of Y on Z and its residual; (d) partial correlation between X and Y adjusting for Z , represented as simple correlation between residuals $X - \hat{x}$ and $Y - \hat{y}$.

There are two explicit formulas for calculating the partial correlation coefficient between X_1 and X_2 after adjusting for X_3 :

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \quad (12.4)$$

and

$$r_{12,3}^2 = \frac{R_{1(2,3)}^2 - R_{1(3)}^2}{1 - R_{1(3)}^2} \quad (12.5)$$

Formula (12.4) can be used recursively to find the partial correlation coefficient between X_1 and X_2 after removing the effects of more than one variable, say X_3 and X_4 as follows. We first find the first order partial correlation coefficients, removing the effects of X_3 , for all three pairs of X_1 , X_2 and X_4 . Then we can use these first order partial correlation coefficients as if they were simple partial correlation coefficients in formula (12.4) to get the second order partial correlation coefficients $r_{12,34}$.

Despite the loss of sign information, (12.5) allows calculation in terms of multiple correlations and provides a meaning to the magnitude of a partial correlation coefficient in a way analogous to R^2 . It says that the squared partial correlation coefficient between X_1 and X_2 adjusting for X_3 is the proportion of variability of X_1 which can be explained jointly by X_2 and X_3 but not singly by X_3 , relative to the total portion that cannot be explained singly by X_3 . By symmetry, we have an alternative but equivalent formula by swapping the roles of X_1 and X_2 in (12.5), and the entailed interpretation. The higher order partial correlation coefficients can be computed in a similar way:

$$r_{12,34}^2 = \frac{R_{1(2,3,4)}^2 - R_{1(3,4)}^2}{1 - R_{1(3,4)}^2} \quad (12.6)$$

Example 12.2 Denote Y by X_4 in Example 12.1. We can find the partial correlation coefficient between X_1 and X_4 after removing the effect of X_2 as follows. Since $r_{14} = 0.7209$, $r_{12} = 0.5317$, $r_{24} = 0.5623$, we have

$$r_{14,2} = \frac{r_{14} - r_{12}r_{24}}{\sqrt{(1 - r_{12}^2)(1 - r_{24}^2)}}$$

$$\begin{aligned}
&= \frac{0.7209 - (0.5317)(0.5623)}{\sqrt{(1 - 0.5317^2)(1 - 0.5623^2)}} \\
&= \frac{0.4219}{0.7004} = 0.6024.
\end{aligned}$$

This means that there is still considerable linear correlation (0.6024) between the numbers in-patient 7 and out-patient after removing the effect of percentage bed-usage. Similarly, $r_{34.2} = 0.6269$, $r_{13.2} = 0.2171$. The second order partial correlations can then be calculated using the ones of the first order, $r_{14.23} = 0.6131$, $r_{34.12} = 0.6367$. These two second order partial correlations show that the out-patient number and the turnaround frequency each still has considerable association with the in-patient number after removing the effects of the two remaining independent variables. However, $r_{24.13} = 0.1364$ is very small, meaning that after removing the effects of in-patient number and turnaround frequency, only $(0.1364)^2 = 0.0186 = 1.86\%$ of the variation of out-patient number is attributed to its dependence on the percentage of bed-usage.

12.2.4 Test of correlation

The significance test for the simple correlation coefficient is given by formula (9.8) of Chap. 9. The test for a zero partial correlation is similar to the t -test for a simple correlation except for an adjustment of the degrees of freedom.

Let the population partial correlation coefficient after removing the effects of q variables be $\rho_{(-q)}$ and the sample counterpart be $r_{(-q)}$. We can test the following hypothesis.

$$H_0: \rho_{(-q)} = 0, \quad H_1: \rho_{(-q)} \neq 0$$

by the t statistic

$$t = \frac{r_{(-q)}\sqrt{n-q-2}}{\sqrt{1-r_{(-q)}^2}}, \quad DF = n - q - 2 \quad (12.7)$$

in the same way as before.

The test for the multiple correlation between Y and X_1, \dots, X_p being zero is the same as the F test in ANOVA that tests the regression of Y on X_1, \dots, X_p .

12.3 Selection of Independent Variables

We have seen in Example 12.1 that not all variables have statistical significance in predicting Y . A natural dictum in model building is to build a parsimonious model without excluding those variables that can contribute significantly to the prediction power of the model. On the other hand, in some studies, the model is meant to be a platform based on which effects of various variables are to be compared and analyzed to get an understanding of a problem, and in such circumstances we may wish to keep variables with p -values as large as 0.15 for a wider scope of the problem and for not inflating the error sum of squares. Therefore, there is no universal approach for variable selection that can fit all purposes of studies. The criteria and algorithms introduced in this section should only be used as tools, not purposes, in applications. The adoption of a method should be coupled with the subject matter knowledge in a particular application. Sometimes, certain otherwise important variables are not included in the model because of their high correlation with some variables already in the model (hence well represented by those variables), or because their ranges of variability are not sufficiently represented in the data set.

12.3.1 Criteria for comparing models

12.3.1.1 Squared multiple correlation R^2

Large value of R^2 corresponds to small value of SS_{Error} and thus the two statistics are equivalent with regard to comparing models. With the same number of variables in the models, the model having the largest R^2 (or the smallest SS_{Error}) is the best. But if we are comparing two models where one is a reduced model of the other, the bigger model always has a larger R^2 and hence we have to judge by experience whether the increase in R^2 is large enough to justify having the additional variables in the bigger model. For example, we deem that a change from 0.8065 to 0.8124 in R^2 does not justify adding a variable to the model to achieve that little improvement. If a formal test is required to decide whether the increase in R^2 is significant enough, we may use the partial F test based on the R^2 for the reduced model with q variables, R_0^2 , and the R^2 for the bigger model with k variables, R_1^2 :

$$F = \frac{n - k - 1}{k - q} \left(\frac{1 - R_0^2}{1 - R_1^2} - 1 \right), \quad (12.8)$$

with $\nu_1 = k - q$, $\nu_2 = n - k - 1$. This is the same test for the null hypothesis that all those regression coefficients which are included in the bigger model but not in the reduced model are zero.

12.3.1.2 Adjusted R^2 or MS_{Error}

The MS_{Error} is the best unbiased estimate of σ^2 if the model is correct and, unlike the SS_{Error} , has taken into account the number of variables in the model. The adjusted coefficient of determination, or adjusted multiple correlation squared is a monotone function of MS_{Error} :

$$\begin{aligned} R_{\text{adj}}^2 &= 1 - \frac{MS_{\text{Error}}}{MS_{\text{Total}}} = 1 - \frac{SS_{\text{Error}}/(n - p - 1)}{SS_{\text{Total}}/(n - 1)} \\ &= 1 - (1 - R^2) \frac{n - 1}{n - p - 1}. \end{aligned} \quad (12.9)$$

The criterion is: the larger the R_{adj}^2 (or the smaller the MSE), the better the model, regardless of the number of variables in the model.

12.3.1.3 Mallows' C_p

The statistic is meant to be an estimate of the number of regression coefficients that should be in the model. Suppose σ^2 is known. If the entertained model with p independent variables is correct, the Error Sum of Square of this model $SS_{\text{Error}}(p)$ has the expected value $(n - p - 1)\sigma^2$. Hence an unbiased estimator of $(p + 1)$, the number of regression coefficients, is

$$\frac{SS_{\text{Error}}(p)}{\sigma^2} - (n - p - 1) + (p + 1) = \frac{SS_{\text{Error}}(p)}{\sigma^2} - n + 2(p + 1).$$

Since σ^2 is not known, some estimate of σ^2 which is not based on the entertained model would be more desirable. A commonly used estimate of σ^2 is the mean square error (MS_{Error}) when all independent variables in the scope of study are put in the model, leading to the following statistic,

$$C_p = \frac{SS_{\text{Error}}(p)}{MS_{\text{Error}}} - n + 2(p + 1). \quad (12.10)$$

Mallows (1973) recommends to focus on the model having the smallest C_p and those that are closest to $(p + 1)$.

12.3.2 Algorithms for variable selection

Once we have adopted a criterion or more, the next step is to have an algorithm for evaluating models.

12.3.2.1 All possible subsets

Compare all $(2^p - 1)$ possible subsets of the variables according to the chosen criterion statistic and selected the optimal ones.

12.3.2.2 Backward elimination method

- (B1) Starting with the full model (include all independent variables in the scope of studies and the intercept).
- (B2) Remove the most insignificant variable in the model according to the t test in the parameter estimates table or in the ANOVA table, i.e. the variable with the largest p -value that is larger than the pre-assigned significance level for removing a variable.
- (B3) Repeat (B2) until no variable in the model with a p -value larger than the significance level for removal.

12.3.2.3 Forward selection method

Use the same test as in Backward Elimination.

- (F1) Starting with the model having no independent variable (i.e. having intercept only).
- (F2) Add the most significant variable into the model, i.e. add the variable with the smallest p -value which is smaller than the pre-assigned significance level for entry.
- (F3) Repeat (F2) until no significant variable can be added.

12.3.2.4 Stepwise selection method

It is a hybrid of the above two methods.

- (S1) Start with the model having no independent variable as in (F1) above.
- (S2) Add a variable into the model as in (F2) above.
- (S3) Remove a variable as in (B2) above.

(S4) Repeat (S2) and (S3) until no variable outside the model can be added and no variable in the model can be eliminated.

Note that not all the methods give the same solution. The researcher should consider the solutions from all methods and see which one(s) make more sense according to the subject matter knowledge, in addition to the criterion.

Whenever feasible, comparing all possible subsets of the independent variables based on the chosen criteria is most desirable. About 30 years ago, textbooks shy away from this method because of the lack of both computing power and efficient computing algorithms to fit a total of $(2^p - 1)$ regression models. Nowadays, the exponentially growing computing power is easily accessible and there are computing algorithms, such as leaps and bounds, that can find the optimal subsets efficiently for a large number of variables. For example, with ten variables and 100 observations, the SAS procedure REG can finish in a split second with a desk-top PC if one wants to see the criteria of R^2 , R^2_{adj} , MS_{Error} and C_p for the five best subsets for each value of p . We therefore recommend using the results of all possible subsets as the road map and see where the Backward Elimination and Stepwise selection solutions stand in the road map before making a decision.

Example 12.3 Consider variables selection for Example 12.1.

Solution The results of all possible subsets are given in Table 12.6.

The largest adjusted R^2 and the smallest C_p criteria clearly choose the model with X_1 and X_3 . The increase in R^2 (from 73.6% to 74.1%) is too small to justify adding X_2 in the model of X_1 and X_3 . Indeed, this increase

Table 12.6 Statistics of all possible subsets.

Number in model	R -square	Adjusted R -square	C_p	Variable in model
1	0.5197	0.4930	13.6697	X_1
1	0.5156	0.4887	13.9206	X_3
1	0.3162	0.2782	26.2368	X_2
2	0.7361	0.7050	2.3034	$X_1 X_3$
2	0.5849	0.5361	11.6384	$X_2 X_3$
2	0.5644	0.5131	12.9098	$X_1 X_2$
3	0.7410	0.6924	4.0000	$X_1 X_2 X_3$

Table 12.7 Summary of backward elimination.

Step	Variable removed	Number vars in	Partial R -square	Model R -square	C_p	F	$\text{Pr} > F$
1	X_2	2	0.0049	0.7361	2.3034	0.30	0.5894

Table 12.8 Summary of stepwise selection.

Step	Variable entered	Variable removed	Number vars in	Partial R -square	Model R -square	C_p	F	$\text{Pr} > F$
1	X_1		1	0.5197	0.5197	13.6697	19.47	0.0003
2	X_3		2	0.2164	0.7361	2.3034	13.94	0.0017

Table 12.9 Estimates in the final model.

Variable	DF	Parameter estimate	Standard error	t	$\text{Pr} > t $	Standardized estimate
Intercept	1	-3369.53992	1571.50137	-2.14	0.0468	0
X_1	1	60.10741	15.95140	3.77	0.0015	0.51390
X_3	1	335.60549	89.89560	3.73	0.0017	0.50915

in R^2 is not statistically significant, as indicated by the F test for X_2 in Table 12.7, which summarizes the Backward Elimination results.

The stepwise selection also gives a model with X_1 and X_3 , as shown in Table 12.8. Note that the forward selection should give the same model from the stepwise selection since there is no removal after each variable enters the model, as shown in Table 12.8. Thus for this example all methods give the same solution. The estimates of the model are given in Table 12.9 and the regression equation is

$$\hat{Y} = -3369.5 + 60.1X_1 + 335.6X_3.$$

12.4 Further Topics in Multiple Regression

Multiple regression has become a vast subject in statistics and is still growing. In this section we briefly introduce some of the related topics. Researchers are referred to specialized books in regression, or statisticians for details.

12.4.1 Model diagnostics and remedies

12.4.1.1 Model assumptions

In multiple regression we assume the conditional distribution of Y given the X variables to be a normal distribution with a conditional mean being a linear function of X variables and a constant conditional variance not depending on X variables. We also assume that n observations from this conditional distribution are independent. We abbreviate the four conditions of linearity, independence, normality and equal variance as the “LINE” condition. The residuals of the regression, $e_i = y_i - \hat{y}_i$, $i = 1, \dots, n$, are the key quantities for checking the LINE condition.

Suppose the subscript of the observations represents the time order of data collection, or an ordering according to some factor like location, the condition of independence may be checked by a plot of the residuals against the subscript. Independence implies that the residuals display a random pattern above and below the zero line when the subscript increases. If there are clearly runs of residuals either above or below the zero line so that there are only a few crossings of zero line by the line that joins the points in the subscript order, we suspect that the independence condition is violated.

If the independence assumption holds, the Durbin–Watson statistic will give

$$DW = \sum_{i=2}^n (e_i - e_{i-1})^2 / \sum_{i=1}^n e_i^2 \approx 2.$$

If $DW \approx 0$, the consecutive pairs of residuals are strongly correlated in the positive direction; if $DW \approx 4$, they are strongly correlated in the negative direction. The independence assumption can also be tested by the non-parametric test for randomness based on runs.

Under the conditions of linearity and constant variance, the scatter plot of residuals against the predicted values would form a horizontal band of uniform width symmetrically around the zero line (see Fig. 12.2(a)). If linearity is violated, the band of points will be curved, instead of being horizontal, and there is not symmetry around the zero line (see Fig. 12.2(b) and (c)). In this case, some nonlinear functions of the X variables should be considered in the model, including quadratic or cubic terms of the X variables. The

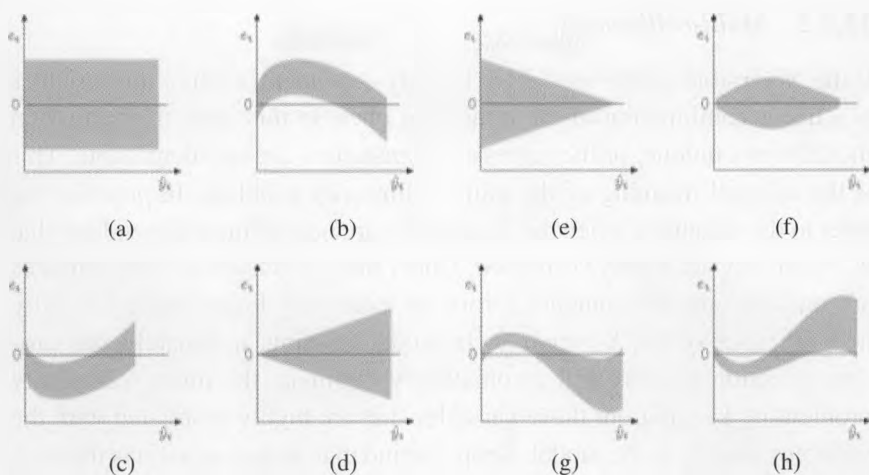


Fig. 12.2 Residual plots for checking assumptions: (a) linearity and constant variance; (b) nonlinear; (c) nonlinear; (d) non-constant variance; (e) non-constant variance; (f) non-constant variance; (g) nonlinear and non-constant variance; (h) nonlinear and non-constant variance.

plots of residuals against individual X variables might indicate which of the X variables needs nonlinear terms in the model.

If the residual plot against the predicted value is basically symmetric about the zero line, but the width of horizontal band changes as \hat{Y} moves from left to right, it is a sign of unequal variance when the values of the X variables change. The most common patterns are the right and left opening megaphone (see Fig. 12.2(d) and (e)), but in some cases, the cluster of points may look like an olive (see Fig. 12.2(f)). Common remedies for non-constant variance are transformation of variables and weighted least-square estimation.

The normality assumption can be checked by the Q-Q normality plot and normality test for the residuals. A common remedy is transformation of variables.

The violations of linearity, constant variance and normality may appear in tandem or together. For example, if the residual plot shows a curved band with non-uniform width, both linearity and constant variance are violated (see Fig. 12.2(g) and (h)). The Box-Cox (1964) method for finding a transform of Y satisfying all three conditions may be useful in the circumstances.

12.4.2 *Multi-collinearity*

If the X variables in the model are linearly dependent, each of the variables is a linear combination of the remaining ones. In this case, the regression model is not unique, or the regression parameters are not identifiable. This is the original meaning of the multi-collinearity problem. In practice, we refer to the situations when the X variables are nearly linear dependent; that is, when they are highly correlated. Under the circumstances, the estimates are unstable and the standard errors are extremely large, leading to false insignificance of the X variables. In model building, fortunately, the variable selection process will automatically eliminate the multi-collinearity problem by keeping out those variables that are highly correlated with the variables already in the model. Keep in mind that in such situations the variables in the model are not necessarily more sensible, according to the subject matter knowledge, than the highly correlated ones outside the model. This is one of the reasons why we have recommended using the results of all possible subsets as a road map for variable selections. In so-called confirmatory analysis where the interest is not on model building, but on testing certain theories within the prescribed framework, we may face the multi-collinearity problem. In such circumstances, we can detect the problem by computing the squared multiple correlation between each of the X variables with the remaining ones. If any of the squared multiple correlations is very large, say larger than 0.9, it is an indication of a multi-collinearity problem. An alternative way is to see if the smallest eigenvalue of the correlation matrix of the X variables is near zero. If indeed there is the multi-collinearity problem, then we have to make a choice between (a) discarding some variables that are statistically insignificant or their proxies, and (b) keeping all X variables in the scope and doing multiple regression on some special linear combinations of all the X variables, which are called "principal components" and will be considered in Chap. 17.

12.4.3 *Outliers and influential observations*

All the problems or issues in the above discussion are concerned with the dependent or independent variables, or both. There are potential problems for the observations too. There could be errors or even blunders in the process of data collection, or contamination in data generation process.

A few observations so obtained may make the model assumptions violated or may be overly influential on the estimates of the parameters. They are called “outliers” and “influential observations” respectively.

There are ways to detect outliers and influential observations and here we only consider some of them. Let $\hat{y}_{(i)}$ be the predicted value using the multiple regression model which has been obtained without the i th observation, called the i th jackknifed predicted value, and $e_{(i)} = y_i - \hat{y}_{(i)}$ the corresponding residual. The i th Studentized jackknifed residual (available as RSTUDENT in the SAS procedures REG and GLM) is $r_{(i)} = e_{(i)} / SE(e_{(i)})$. As a rule of thumb, the i th observation is an outlier with respect to the Y values if $|r_{(i)}| > 3$, not an outlier if $|r_{(i)}| < 2$, and undetermined case if $2 < |r_{(i)}| \leq 3$. The “leverage” of the i th observation, or the i th leverage is defined as $h_{ii} = x_i^T (X'X)^{-1} x_i$, where x_i is the i th row of the matrix X , with the property that $\sum_{i=1}^n h_{ii} = p + 1$. The i th observation is considered as an outlier with respect to the independent variables if $h_{ii} > 3(p + 1)/n$, not an outlier if $h_{ii} < 2(p + 1)/n$, and undetermined if it is in between. Cook and Weisberg (1982) propose measuring the influence of the i th observation on estimation of the regression coefficients by the displacement of the estimates when the i th observation is not used. The statistic, called Cook's D , is defined as

$$D_i = (\hat{\beta} - \hat{\beta}_{(i)})^T (X^T X)^{-1} (\hat{\beta} - \hat{\beta}_{(i)}) / (p + 1) MS_{\text{Error}}, \quad (12.11)$$

where $\hat{\beta}_{(i)}$ is the vector of estimates without the i th observation. The distribution is close to $F(p + 1, n - p - 1)$. As a rule of thumb, we declare the i th observation as influential if D_i is larger than the 50th percentile of $F(p + 1, n - p - 1)$, not influential if smaller than the 20th percentile, undetermined if in between.

12.4.4 Interaction effects

When the optimal linear combination of independent variables cannot fully explain the variability of Y , we might need interaction terms in the model. In a multiple regression, the interaction effects of two variables say X_1 and X_2 , are usually indicated by their product $X_1 X_2$. We may treat the product terms as ordinary variables in the multiple regression.

In judging whether the interaction effects should be considered in a model, the subject matter knowledge should play a primary role. When

Table 12.10 Percentage of positive residuals, close to 50%.

X_1	X_2		
	Low	Middle	High
Low	40	51	50
Middle	50	48	51
High	51	50	49

Table 12.11 Percentage of positive residuals, varying from row to row.

X_1	X_2		
	Low	Middle	High
Low	20	30	40
Middle	35	50	68
High	52	65	80

such knowledge is not available, we may first consider a model without interaction and examine the residual plots against individual X variables to see whether the linearity condition is violated. If so, higher order terms, including the cross products, may be needed.

The following is also a useful way to detect the need for interaction terms between two continuous variables after we have done a multiple regression. Divide the values of each variable into three groups, Low, Middle and High. Calculate the percentage of positive residuals for each of the $3 \times 3 = 9$ cells of the cross-classified two-way table of the value groups of X_1 and X_2 . For examples, in Table 12.10, all percentages are around 50%; there is no need for interaction effects between X_1 and X_2 in the model; but in Table 12.11, the percentages are quite different row to row, suggesting that we should add the cross-product terms such as $X_1 X_2$, $X_1^2 X_2$, or $X_1 X_2^2$.

12.4.5 Dummy variables for groups of data

Very often we have data for both genders and wonder whether we can pool them together in doing multiple regression. If there is no gender difference, pooling the data will yield better estimates of the parameters. And if there

is a difference, we want to estimate the difference with standard error and confidence limits. These two questions cannot be answered by separate regression analyses for the genders. A “dummy variable”, which equals to either zero or one, may be used if we want to study two groups of data simultaneously. A dummy variable is also called an indicator variable. Let Z be a dummy variable so that $Z = 1$ if the observation is from the first group and $Z = 0$ if from the second group. We create p more variables, $X_i Z, i = 1, \dots, p$, which are the products of X_i and Z and also called interaction effects between X_i and Z as in previous section, and then we may have a regression equation

$$\begin{aligned} \hat{Y} = & a + bZ + b_1X_1 + b_2X_2 + \dots + b_pX_p + b_{p+1}X_1Z \\ & + b_{p+2}X_2Z + \dots + b_{p+p}X_pZ \end{aligned} \quad (12.12)$$

for the model of

$$\begin{aligned} \mu_{Y|X_1, \dots, X_p} = & \alpha + \beta Z + \beta_1X_1 + \beta_2X_2 + \dots + \beta_pX_p + \beta_{p+1}X_1Z \\ & + \beta_{p+2}X_2Z + \dots + \beta_{p+p}X_pZ. \end{aligned} \quad (12.13)$$

To answer the above two questions, we need to test

$$\begin{aligned} H_0: & \beta = \beta_{p+1} = \beta_{p+2} = \dots = \beta_{p+p} = 0; \\ H_1: & \text{The full model corresponding to (12.13) is true.} \end{aligned} \quad (12.14)$$

If H_0 is not rejected, the two sets of data can be pooled to get a unique regression equation; otherwise, they cannot be pooled, and two separate regression equations are needed.

However, in practice, we often test the above hypotheses by two steps.

Firstly, we test (note that β is absent)

$$\begin{aligned} H_0: & \beta_{p+1} = \beta_{p+2} = \dots = \beta_{2p} = 0; \\ H_1: & \text{The full model corresponding to (12.13) is true.} \end{aligned} \quad (12.15)$$

This is the same as the “parallelism assumption” in analysis of covariance where the primary interest is comparing the group means after adjusting for the (common) effects of covariates (the independent variables in our context), and can also be regarded as testing the interaction effects between

X_i and Z in (12.13). The test statistic is

$$F = \frac{(SS_{\text{Error}}(H_0) - SS_{\text{Error}}(H_1))/p}{SS_{\text{Error}}(H_1)/(n - 2(p + 1))}, \quad (12.16)$$

where $SS_{\text{Error}}(H_0)$ and $SS_{\text{Error}}(H_1)$ are the SS for error after fitting models under H_0 and H_1 respectively.

If $F < F_\alpha(p, n - 2(p + 1))$, the parallelism condition holds (all the coefficients corresponding to interaction effects in model (12.13) are zero), and the conditional full model turns to be

$$\mu_{Y|X_1, \dots, X_p} = \alpha + \beta Z + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p. \quad (12.17)$$

Accordingly the regression equation will be

$$\hat{Y} = a + bZ + b_1 X_1 + b_2 X_2 + \dots + b_p X_p. \quad (12.18)$$

Now, as the higher order interaction term (say, $X_i Z$ s) is statistically insignificant, we can test the main effect term (say, Z or X_i s). Then as the second step, under the condition of parallelism we can further test

$$H_0: \beta = 0;$$

$$H_1: \text{The full model corresponding to (12.17) is true.} \quad (12.19)$$

The test statistic is

$$F = \frac{(SS_{\text{Error}}(H_0) - SS_{\text{Error}}(H_1))/p}{SS_{\text{Error}}(H_1)/(n - p - 2)} \quad (12.20)$$

which has a distribution of $F(p, n - p - 2)$ when H_0 is true.

If $F < F_\alpha(p, n - p - 2)$, we can pool the two groups to get a unique regression equation in the form of (12.21).

$$\mu_{Y|X_1, \dots, X_p} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p. \quad (12.21)$$

Otherwise, If $F \geq F_\alpha(p, n - p - 2)$, under the parallelism condition in the preceding paragraph, the difference in group means can be estimated within the model (12.17). The sample coefficient of Z estimates the difference between the mean of the first group and that of the second group.

For k groups of data, we need $k-1$ dummy variables, say $Z_j = 1$ for the j th group, and 0 otherwise. Note that the k th group (the last group) is identified by $Z_j = 0$ for $j = 1, \dots, k-1$. So we don't need k dummy variables. We next form the $p(k-1)$ product terms, $X_i Z_j, i = 1, \dots, p$

and $j = 1, \dots, k - 1$. The coefficient of Z_j is the increase in intercept of the j th group relative to the k th group, and the coefficient of $X_i Z_j$ is the increase in the coefficient of X_i for the j th group relative to the k th group. Thus if there is a natural reference group, it is more convenient to label it as the last group. The estimates and tests are analogous to the case of two groups.

12.5 Path Analysis

In some applications, we have good reasons to think of causal relations among the variables in a more complex structure than just one dependent variable and the remaining ones being independent variables. In the circumstances, we need to make use of subject matter knowledge to draw a so-called path diagram to delineate the hypothetical causal structure among the variables concerned. We first consider an example.

Example 12.4 In a study about births (Shao-Xian Wang, 1983), the researchers randomly took a sample of 2,000 cases from 11,309 married women in west Beijing, who have participated in the study. The following six variables were considered: Age in years (X_1), education ($X_2 = 0$ for illiterate, 1 for primary school, etc.), age in years when first married (X_3), number of pregnancy ($X_4 = 1, 2$, etc.), live birth rate in percentage (X_5) and number of live births ($Y = 1, 2$, etc.). The aim is to study the possible causal structure among the X variables and their effects on Y .

The first thought would be a multiple regression of Y on all X variables, all in standardized form. The regression equation is:

$$\hat{Y} = 0.3415x_1 - 0.0941x_2 - 0.1921x_3 + 0.4585x_4 + 0.2688x_5.$$

The coefficients are standardized so that direct comparisons are legitimate. The direct effects of the X variables are represented by the coefficients in a straightforward way. For instances, the higher the education level or the older the age of first marriage, the less live births; the larger the other three variables, the more live births. Education has very small direct effects.

We anticipate that education has effects on the age of first marriage and on the number of pregnancy, and hence should have indirect effect through these two intermediate variables. A path analysis based on the so-called structurally related equations is beyond the scope of this book. Here we

illustrate the simple approach of effects accounting based on regression and correlation analysis.

Step 1 Draw a path diagram

Draw an initial path diagram with the aid of professional knowledge about the nature of variables, progression order of outcomes of variables, etc., according to the following rules:

- An arrow is drawn to a “resultant variable” (or endogenous variable) from each of its “causal variables” (or exogenous variables).
- To each resultant variable there is also an arrow originated from an unobservable error variable (or residual variable) which represents all those possible factors that may have effects on the resultant variable but not identified.
- A double-headed arrow connects each pair of causal variables that are thought to be correlated but not resultant to each other.

The initial path diagram for Example 12.4 is given in Fig. 12.3. There are two purely causal variables, age (X_1) and education (X_2), which are not resultant to each other. The number of live births (Y) is the only purely resultant variable that is not viewed as causal variable to any other variable. The age of first marriage (X_3), pregnancy number (X_4) and live birth rate (X_5) are intermediate variables that are both causal and resultant variables.

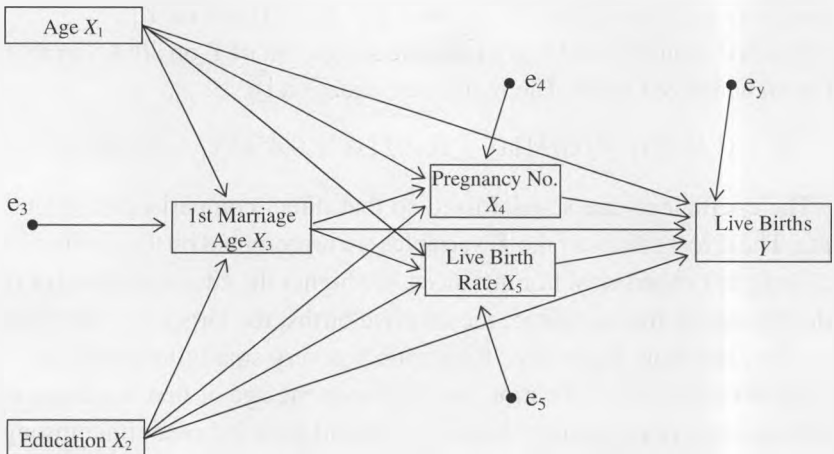


Fig. 12.3 Initial path diagram for live birth study.

For example, X_3 is a resultant variable of X_1 and X_2 plus error e_3 , but also a causal variable to Y , while X_4 is a resultant variable of X_1 , X_2 and X_3 plus an error e_4 , but also a causal variable of Y , etc.

Step 2 Find the path coefficients

The path coefficients for the paths to a resultant variable from all its causal variables are the respective standardized regression coefficients that are statistically significant in a multiple regression of the resultant variable on its causal variables. For the path to the resultant variable from its error variable we show $1 - R^2$, in rounded percentage, to indicate the proportion of the variability that is not accountable by the existence of the causal variables. For a double-headed arrow, the path coefficient is the simple correlation coefficient.

For Example 12.4, we have to fit the following four multiple regressions in standardized form, where P_{jk} are the standardized regression coefficients, preferably all statistically significant:

$$X_3 = P_{31}X_1 + P_{32}X_2 + e_3,$$

$$X_4 = P_{41}X_1 + P_{42}X_2 + P_{43}X_3 + e_4,$$

$$X_5 = P_{51}X_1 + P_{52}X_2 + P_{53}X_3 + e_5,$$

$$Y = P_{Y1}X_1 + P_{Y2}X_2 + P_{Y3}X_3 + P_{Y4}X_4 + P_{Y5}X_5 + e_Y.$$

Using Backward Elimination with a significance level at $\alpha = 0.05$, we obtain the following standardized regression coefficients together with the coefficient of multiple determination for each regression equation:

$$X_3 = 0.6095X_2 (R_3^2 = 0.3715)$$

$$X_4 = 0.6719X_1 - 0.0711X_2 - 0.2622X_3 (R_4^2 = 0.5014)$$

$$X_5 = 0.4608X_1 - 0.0325X_2 (R_5^2 = 0.1923)$$

$$Y = 0.3415X_1 - 0.0941X_2 - 0.1921X_3 + 0.4585X_4 \\ + 0.2688X_5 (R_Y^2 = 0.8298)$$

The path coefficients are put in place as shown in Fig. 12.4.

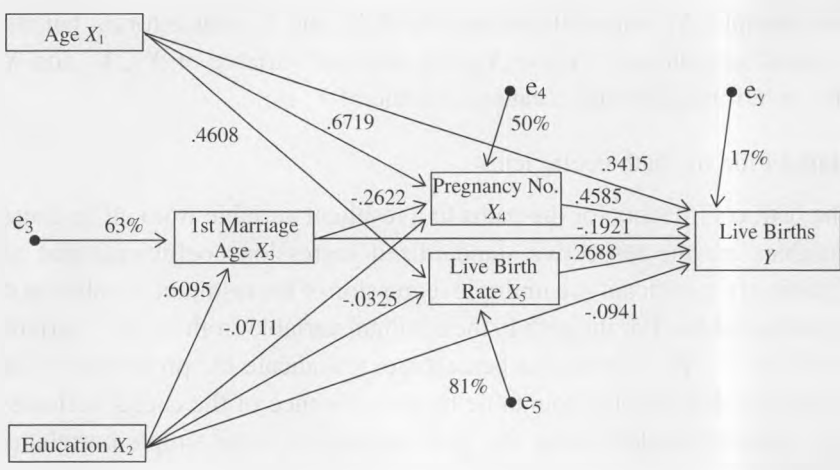


Fig. 12.4 Statistically significant paths.

Table 12.12 Effects of independent variables through all possible routes to Y .

Causal variables	Direct effects	Indirect effect through intermediate variables						Subtotal	Total
		X_3	X_4	X_5	X_3, X_4	X_3, X_5			
X_1 Age	0.3415	0	0.3081	0.1239	0	0	0.4320	0.7735	
X_2 Education	-0.0941	-0.1171	-0.0326	-0.0087	-0.0733	0	-0.2317	-0.3258	
X_3 1st Marriage Age	-0.1921	NA	-0.1202	0	NA	NA	-0.1202	-0.3123	
X_4 Pregnancy No.	0.4585	NA	NA	NA	NA	NA	NA	0.4585	
X_5 Live Birth Rate	0.2688	NA	NA	NA	NA	NA	NA	0.2688	

Step 3 Construct the effects table

The direct and indirect effects of the independent variables on the dependent variables through all possible routes to Y are summarized in Table 12.12. Those routes that are not included in the original path diagram (Fig. 12.3) are indicated by NA (for not appropriate/available). The routes that have been proposed but later deleted because of statistical insignificance will

be given a zero value. For example, the path from X_1 to X_3 is canceled; hence the indirect effects of X_1 via X_3 , via X_3 and X_4 and via X_3 and X_5 are all zero. The standardized regression coefficients in the regression equation for Y measure the direct effects of all independent variables. The indirect effect of an independent variable through intermediate variables is represented by the product of the corresponding path coefficients along the route to Y . For example, the indirect effect of education X_2 on Y through X_3 is $0.6095 \times (-0.1921) = -0.1171$, and through X_3 and X_4 is $0.6095 \times (-0.2622) \times 0.4585 = -0.0733$.

Step 4 Interpret the results

- (a) Pregnancy number (X_4) has the largest immediate effect (0.4585), and second in total.
- (b) Age (X_1) has the second largest immediate effect (0.3415) but the largest indirect effect (0.4320), giving a total (0.7735) larger than pregnancy number (X_4).
- (c) Education (X_2) has the smallest direct effect (-0.0941 , ignoring the sign) but second largest indirect effect (-0.2317).
- (d) Directly or indirectly, education (X_2) and age of first marriage (X_3) have negative effects, while the other three variables have positive effects.
- (e) The total effects of education (-0.3258) and age of first marriage (-0.3123) are about the same, the former having less direct effect but more indirect effect.
- (f) For indirect effects through first marriage age (X_3), education (X_2 with -0.1171) is more influential than age (X_1 with 0). But for indirect effects through X_4 or X_5 , age is more influential than education (0.3081 versus -0.0326 and 0.1239 versus -0.0087).
- (g) Education has slightly larger effect than age through the composite route of X_3 and X_4 (-0.0733 versus 0). If we look at the subtotal, age has the largest indirect effect (0.4320); the next is education (-0.2317) and then age of first marriage (-0.1202).
- (h) Given the large proportions of the variability due to residual variables of education (63%), pregnancy number (50%) and live birth rate (81%), there may be important causal variables that have not been identified in the study.

12.6 Computerized Experiments

Experiment 12.1 Computation for Example 12.1 Assume that the data of year, Y , X_1 , X_2 , X_3 are stored in the text file named “ex14-1.dat” in a floppy disk. The results contained in Tables 12.2–12.8 can be reproduced by the following SAS codes:

All keywords in the program are in capital letters for easy reference, but not essential. In program 12.1, line 02 points to the location of the data file and line 03 reads the data. The first 3 lines create the dataset named “inpatients” with five variables and 20 observations. The first MODEL statement additionally requests standardized regression coefficients (STB), confidence limits for the mean (CLM) and individual predicted values (CLI), residuals and their standard errors, Studentized residuals and Cook’s D. The second MODEL statement requests statistics of criteria R^2 , R^2_{adj} , MS_{Error} and C_p for all possible subsets regression. The third MODEL statement is for Backward Elimination with $\alpha = 0.05$ and the last MODEL statement is for stepwise selection with $\alpha = 0.10$ to enter a variable and $\alpha = 0.05$ for removing a variable.

Experiment 12.2 Impact of normality Simulate X_1 as $N(10, 3^2)$, X_2 as the square of a variable from $N(10, 3^2)$, X_3 as the cube of a variable from $N(10, 3^2)$ and E from $N(0, 1)$. Then defined $Y_1 = X_1 + X_2 + X_3 + E$ and $Y_2 = X_1 + X_2 + X_3 + E^2$. Generate a sample of size 30 for the variables $(X_1, X_2, X_3, E, Y_1, Y_2)$. Separately carry out regression of Y_1 and

Program 12.1 Computation for Example 12.1.

Line	Program
01	DATA inpatients;
02	INFILE 'a: \ex14-1.dat';
03	INPUT year y x1-x3;
04	PROC REG;
05	MODEL y= x1-x3 / STB CLM CLI R;
06	MODEL y= x1-x3 / SELECTION= RSQUARE ADJR SQ MSE
07	MODEL y= x1-x3 / SELECTION=B SLS=0.05';
08	MODEL y= x1-x3 / SELECTION=STEPWISE SLE=0.1'
09	QUIT;

Program 12.2 Impact of normality.

Line	Program	Line	Program
01	DATA A;	11	PROC REG;
02	DO I=1 TO 30;	12	MODEL Y1=X1 X2 X3/R;
03	X1=RANNOR(0)*3+10;	13	OUTPUT OUT=REG1 R=RESID1;
04	X2=(RANNOR(0)*3+10)**2	14	MODEL Y2=X1 X2 X/R;
05	X3=(RANNOR(0)*3+10)**3	15	OUTPUT OUT=REG2 R=RESID2;
06	E=RANNOR(0);	16	DATA REG;
07	Y1=E+X1+X2+X3;	17	SET REG1 REG2;
08	Y2=E**2+X1+X2+X3;	18	PROC UNIVARIATE NORMAL;
09	OUTPUT;	19	VAR RESID1 RESID2;
10	END;	20	RUN;

Program 12.3 Impact of collinearity.

Line	Program	Line	Program
01	DATA A;	07	END;
02	DO I=1 TO 30;	08	PROC REG;
03	X1=RANNOR(0)*3+10;	09	MODEL Y=X1 X2;
04	X2=RANNOR(0)+X1*2;	10	MODEL Y=X1 X2/
05	Y=RANNOR(0)+X1*3+1		SELECTION=FORWARD;
06	OUTPUT;	11	RUN;
		12	QUIT;

Y_2 on X_1 , X_2 and X_3 . Compare the regression coefficients with the true values (1, 1, 1). Use the residuals to check the model assumptions. In the SAS codes below, lines 01–10 create the data, lines 11–15 are for regression analysis and output the residuals to dataset REG1 and REG2, lines 16 and 17 merge the two datasets into one containing both sets of residuals, and lines 18–20 test the normality of both sets of residuals.

Experiment 12.3 Collinearity Program 12.3 simulates a sample of size 30 for the variables (X_1 , X_2 , Y) and carry out regressions analysis for Y on X_1 and X_2 and forward selection of variables, where X_1 is from $N(10, 3^2)$, $X_2 = 2X_1 + e$, e from $N(0, 1)$, and Y from $N(3X_1 + 1, 1)$. Repeat the experiment 20 times. Compare the parameter estimates and variable selection results for the 20 samples and discuss how does the correlation between X_1 and X_2 affect the results.

Table 12.13 Data on manpower and workload for 17 hospitals.

i	X_1	X_2	X_3	X_4	X_5	Y
1	15.57	2463	472.92	18.0	4.45	566.52
2	44.02	2048	1339.75	9.5	6.92	596.82
3	20.42	3940	620.25	12.8	4.28	1033.15
4	18.74	6505	568.33	36.7	3.90	1603.62
5	49.20	5723	1497.60	35.7	5.50	1611.37
6	44.92	11520	1365.83	24.0	4.60	1613.27
7	55.48	5779	1687.00	43.3	5.63	1854.17
8	59.28	5969	1639.92	46.7	5.15	2160.55
9	94.39	8461	2872.33	78.7	6.18	230.58
10	128.02	20106	3655.08	180.5	6.15	3505.93
11	96.00	13313	2912.00	60.9	5.88	3571.89
12	131.42	10771	3921.00	103.7	4.88	3741.40
13	127.21	15543	3865.67	126.8	5.50	4026.52
14	252.90	36194	7684.10	157.7	7.00	10343.81
15	409.20	34703	12446.33	169.4	10.78	11732.17
16	463.70	39204	14098.40	331.4	10.78	15414.94
17	510.21	86533	15524.00	371.6	6.35	18854.45

12.7 Practice and Experiments

1. Yu and Xiang (1993) collected data on the following variables of manpower and workload for 17 hospitals (see Table 12.13)

X_1 = Number of in-patients per day

X_2 = Number of patients X-rayed per month

X_3 = Occupied bed-days per month

X_4 = Population size in thousands targeted for service

X_5 = Number of days in hospital per in-patient

Y = Man-hour deployed per month.

- (1) Find the regression equation for Y on all five X variables and report the ANOVA result.
- (2) Use all variable selection methods and see whether the results are the same. Make a recommendation based on the results.
- (3) For the model you have recommended in (2), present the ANOVA, estimates and R^2 . Are there any outliers or observations, which are overly influential?
- (4) If there are outliers and overly influential observations in (3), delete them and do the regression again according to the model in (3).

Table 12.14 Data of a hospital during 1976–1986.

Year	Y	X_1	X_2
1976	41.56	2.704	30.89
1977	47.68	2.888	32.20
1978	53.06	2.746	36.87
1979	73.68	2.813	39.09
1980	125.35	3.072	44.14
1981	167.53	3.310	48.11
1982	215.21	4.005	51.41
1983	244.55	4.166	52.93
1984	315.32	4.429	59.61
1985	413.07	4.862	67.00
1986	425.27	5.176	68.98

2. Wang (1993) reported the data of a certain hospital during 1976–1986 for the following variables: (see Table 12.14)

Y = Income in thousand dollars (RMB)

X_1 = Number of discharged patients in thousands

X_2 = Actually occupied bed-days (in thousands)

- (1) Find the regression equation of Y on X_1 and X_2 and report the analysis of variance results.
 - (2) Find the partial correlation coefficient between X_1 and Y removing the effect of X_2 .
3. Perform a regression analysis for mathematics score (X) on language score (Y) and IQ (Z) in Problem 5 of Chap. 9. Find the partial correlation coefficient between X and Y controlling the effects of Z . Comparing the partial correlation coefficient with the simple correlation coefficient, what do you see?
4. For Example 12.1, add the variable $X_4 = X_2X_3$ and carry out a regression of Y on X_1, X_2, X_3 and X_4 and test whether there is statistically significant interaction effect between X_2 and X_3 .
5. Carry out a path analysis in Problem 1, only for the variables X_1, X_2, X_5 and Y .

(1st edn. Kai Ng, Tong Wang, Jiqian Fang; 2nd edn. Jinxin Zhang, Jiqian Fang)



Chapter 13

Measures of Multi-variate Data and Multi-variate Analysis of Variance

The observations in a medical research usually include three types of variables: variable for group, response variables and covariates. In most cases, especially in clinical studies, the observations may include more than one response variable. For example, when measuring the blood pressure of a patient both systolic pressure and diastolic pressure are noted; the growth status and physical development of a child are usually measured with more than ten or 20 variables, such as height, weight, head circumference, chest circumference, etc. In such circumstances, although ordinary statistics such as mean, standard deviation, and standard error can be used for description purposes, and statistical inference such as hypothesis testing can be carried out for each variable separately, there are substantial disadvantages as discussed below:

1. Overall information of multi-variable cannot be fully used.
2. No integral conclusion can be made when the results of the hypothesis testing for each variable disagrees with each other.
3. The relationship among variables cannot be examined.

To avoid the drawbacks of applying uni-variable analysis to multi-variable data, multi-variate statistical procedures should be used.

13.1 Multi-variate Statistical Description

Example 13.1 Five patients with high serum lipid were treated with a drug, and the observations after the treatment are listed in Table 13.1. We will use these data to demonstrate how to calculate multi-variate statistics. (Note: A sample with five observations is usually not large enough for

Table 13.1 Lowering blood lipid after treatment.

Subject (j)	Decrease of cholesterol (mg %)	Decrease of triglyceride (mg %)
1	16	-4
2	21	46
3	57	-40
4	-14	107
5	17	86
\bar{x}	18.2	39.0
s^2	744.7	3743.0

scientific research; this can only be used as a hypothetical example to illustrate the method of obtaining multi-variate statistics.)

13.1.1 Mean vector

There are two response variables, X_1 and X_2 , the observations for the first subject ($j = 1$) in Table 13.1 can be described by a two-dimensional vector

$$\mathbf{X}_1 = \begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix} = (X_{11} \ X_{21})' = (16 \ -4)'.$$

When $j = 2$, the observations can be written as

$$\mathbf{X}_2 = \begin{pmatrix} X_{12} \\ X_{22} \end{pmatrix} = (X_{12} \ X_{22})' = (21 \ 46)'.$$

In general, the j th observation can be described as

$$\mathbf{X}_j = \begin{pmatrix} X_{1j} \\ X_{2j} \end{pmatrix} = (X_{1j} \ X_{2j})'.$$

In this example, the mean of X_1 is 18.2, and that of X_2 is 39.0, which can also be described by a mean vector:

$$\bar{\mathbf{X}} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \end{pmatrix} = (\bar{X}_1 \ \bar{X}_2)' = (18.2 \ 39.0)'.$$

More generally, if there are n observations, and each of them has p response variables X_1, X_2, \dots, X_p , the observed data of the first subject is $X_{11}, X_{21}, \dots, X_{p1}$ and can be described by a vector of p dimensions as $\mathbf{X}_1 = (X_{11} \ X_{21} \ \dots \ X_{p1})'$, and that of the j th subject having measurements $X_{1j}, X_{2j}, \dots, X_{pj}$ can be presented using a vector of p dimensions

$$\mathbf{X}_j = (X_{1j} \ X_{2j} \ \dots \ X_{pj})'.$$

Consequently, the means of p variables for all observed subjects is a vector with p dimensions:

$$\bar{\mathbf{X}} = (\bar{X}_1 \ \bar{X}_2 \ \dots \ \bar{X}_p). \quad (13.1)$$

The sample mean of i th variable is

$$\bar{X}_i = \frac{1}{n} \sum_{j=1}^n X_{ij}.$$

13.1.2 Covariance matrix

The variances of response variables X_1 and X_2 in this example are $S_{11} = S_1^2 = 744.7$ and $S_{22} = S_2^2 = 3743.0$ respectively, and the covariance of X_1 and X_2 is

$$\begin{aligned} S_{12} = S_{21} &= \frac{1}{n-1} \sum_{j=1}^n (X_{1j} - \bar{X}_1)(X_{2j} - \bar{X}_2) \\ &= \frac{1}{5-1} [(16-18.2)(-4-39) + \dots + (17-18.2)(86-39)] \\ &= -1401.25. \end{aligned}$$

The variances and the covariance of two variables are the elements in the following 2×2 matrix

$$\mathbf{S}^2 = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} 744.7 & -1401.25 \\ -1401.25 & 3743.0 \end{pmatrix}$$

which is called a variance-covariance matrix or covariance matrix in brief.

Generally, if there are p response variables X_1, X_2, \dots, X_p , the variance-covariance matrix of the sample is a $p \times p$ matrix, labeled as

$$S^2 = \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{pmatrix}. \quad (13.2)$$

The diagonal of the matrix consists of the variance of each variable

$$S_{ii} = S_i^2 = \frac{1}{n-1} \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2 \quad i = 1, 2, \dots, p.$$

The elements on each side of the diagonal are the covariance between variables.

$$S_{ik} = S_{ki} = \frac{1}{n-1} \sum_{j=1}^n (X_{ij} - \bar{X}_i)(X_{kj} - \bar{X}_k) \quad \begin{matrix} i \neq k \\ i, k = 1, 2, \dots, p \end{matrix}$$

S is a symmetric matrix if $S_{ik} = S_{ki}$.

13.1.3 Correlation matrix

The correlation coefficient of X_1 and X_2 in this example is

$$r_{12} = r_{21} = \frac{S_{12}}{\sqrt{S_{11}S_{22}}} = -0.8393.$$

The correlation coefficient between any variable and itself is 1, that is $r_{11} = r_{22} = 1$. All correlation coefficients are jointly expressed as in the following 2×2 matrix, called a correlation matrix.

$$\mathbf{R} = \begin{pmatrix} 1 & -0.8393 \\ -0.8393 & 1 \end{pmatrix}.$$

Generally, if there are p response variables X_1, X_2, \dots, X_p , we have a $p \times p$ correlation matrix by summarizing all the correlation coefficients in

one matrix.

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix}, \quad (13.3)$$

$$r_{ik} = r_{ki} = \frac{S_{ik}}{\sqrt{S_{ii}S_{kk}}} \quad i \neq k \\ i, k = 1, 2, \dots, p$$

where r_{ik} ($i \neq k$) is the sample correlation coefficient of X_i and X_k . \mathbf{R} is also a symmetric matrix if $r_{ik} = r_{ki}$.

13.1.4 Multi-variate normal distribution

Let $\boldsymbol{\mu}$ represent the population mean vector which can be expressed as follows, corresponding to the sample mean vector shown in Eq. (13.1).

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = (\mu_1 \quad \mu_2 \quad \cdots \quad \mu_p)'. \quad (13.4)$$

Let $\boldsymbol{\Sigma}^2$ be the population variance-covariance which can be expressed as a $p \times p$ covariance matrix, corresponding to the sample covariance matrix showed in Eq. (13.2).

$$\boldsymbol{\Sigma}^2 = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}, \quad (13.5)$$

where σ_{ii} is the population variance of the i th response variable and σ_{ik} ($i \neq k$) the population covariance of the i th and k th response variables. We assume in this chapter that n observed vectors of a sample of \mathbf{X} following a p -dimensional multi-variate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^2)$, whose

density function is

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2} |\Sigma^2|^{1/2}} \exp[-(\mathbf{X} - \mu)' \Sigma^{-2} (\mathbf{X} - \mu)/2], \quad (13.6)$$

where $|\Sigma^2|$ is the determinant of matrix Σ^2 and Σ^{-2} is the inverse of Σ^2 . It is easy to verify that Eq. (13.6) is just the density function of a univariate normal distribution when $p = 1$.

13.2 Comparison between Two Mean Vectors — Hotelling's T^2 Test

13.2.1 Test for single mean vector $\mu = \mu_0$

We are familiar with univariate hypothesis testing under the assumption that observed variable X follows a normal distribution $N(\mu, \sigma^2)$. To test the null hypothesis $H_0 : \mu = \mu_0$ based on the sample mean \bar{X} , the appropriate test statistic is t

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

The t -statistic can also be expressed as

$$T = \sqrt{n}(\bar{X} - \mu_0)S^{-1}. \quad (13.7)$$

Taking squares on both sides of Eq. (13.7) and making slight changes, we have

$$T^2 = n(\bar{X} - \mu_0)S^{-2}(\bar{X} - \mu_0). \quad (13.8)$$

When there are more than one response variable, we can replace \bar{X} with sample mean vector $\bar{\mathbf{X}}$, and sample variance S^2 with sample variance-covariance matrix \mathbf{S}^2 , then T^2 is generalized as Hotelling's T^2 . That is

$$T^2 = n(\bar{\mathbf{X}} - \mu_0)' \mathbf{S}^{-2} (\bar{\mathbf{X}} - \mu_0), \quad (13.9)$$

where \mathbf{S}^2 is a matrix and \mathbf{S}^{-2} is its inverse matrix.

When there is only one response variable, i.e. $p = 1$ and $H_0 : \mu = \mu_0$ is true, we have $F = t^2$. As an extension, the relationship of F and

Hotelling's T^2 is

$$F = \frac{n-p}{(n-1)p} T^2, \quad \begin{matrix} \nu_1 = p \\ \nu_2 = n-p \end{matrix} \quad (13.10)$$

which can be used to test the null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ based on mean vector $\bar{\mathbf{X}}$ and then compared with critical values of F . When n is large enough, F approximately follows a χ^2 distribution with p degrees of freedom.

Example 13.2 Conduct a statistical inference of whether the drug has the effect on lowering blood lipid with the data showed in Table 13.1.

Solution We know from the data that there are two response variables presenting the effect of blood lipid changes. If the mean vector in population is not equal to $\boldsymbol{\mu}_0 = (0 \ 0)$, we can conclude that the drug significantly affect the change of blood lipid.

$$H_0 : \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad H_1 : \boldsymbol{\mu} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

By calculation,

$$n = 5, \quad p = 2, \quad \bar{\mathbf{X}} = \begin{pmatrix} 18.2 \\ 39.0 \end{pmatrix},$$

$$\mathbf{S}^2 = \begin{pmatrix} 744.7 & -1401.25 \\ -1401.25 & 3743.0 \end{pmatrix}, \quad \mathbf{S}^{-2} = \begin{pmatrix} 0.0045 & 0.0017 \\ 0.0017 & 0.0009 \end{pmatrix}.$$

Compute Hotelling's T^2 with Eq. (13.9) and the value of F in (13.10):

$$\begin{aligned} \mathbf{T}^2 &= n \bar{\mathbf{X}}' \mathbf{S}^{-2} \bar{\mathbf{X}} = 5(18.2 \ 39.0) \begin{pmatrix} 0.0045 & 0.0017 \\ 0.0017 & 0.0009 \end{pmatrix} \begin{pmatrix} 18.2 \\ 39.0 \end{pmatrix} \\ &= 26.4697, \end{aligned}$$

$$F = \frac{n-p}{(n-1)p} T^2 = \frac{5-2}{(5-1) \times 2} (26.4697) = 9.9261,$$

$$\nu_1 = 2, \quad \nu_2 = 3.$$

For $F = 9.9261$, $P = 0.0476$, $H_0 : \boldsymbol{\mu}_0 = (0 \ 0)'$ should be rejected at the significant level 0.05. If any element of $\boldsymbol{\mu}$ (the mean of changes) is different

from 0, then we can say that the drug is effective in lowering blood lipid. However, if we work out two times of univariate t test $\mu_1 = 0$ and $\mu_2 = 0$ respectively, we will have $t = 1.4913, \nu = 5 - 1 = 4, P = 0.2101$ for X_1 , and $t = 1.4254, \nu = 5 - 1 = 4, P = 0.2272$ for X_2 , the null hypothesis could not be rejected. This example shows that Hotelling's T^2 test is more powerful than univariate t test.

13.2.2 Test for two mean vectors $\mu_1 = \mu_2$

Example 13.3 A sample of 18-year-old men in a country of Shaanxi xi province, China was randomly selected and their heights, weights and chest circumferences were measured. The results are listed in Table 13.2. Find out whether the difference between the measurements of 18-year-old men in rural area and urban area is statistically significant.

It is assumed that the observations of a variable, say height, in the two groups were distributed according to univariate normal distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ respectively. To test $H_0 : \mu_1 = \mu_2$ using the sample means of \bar{X}_1 and \bar{X}_2 , we use the statistic

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{n_1 + n_2}{n_1 n_2} S_c^2}}. \tag{13.11}$$

Table 13.2 Physical measurements of 18-year-old men in a county of Shaanxi province.

Rural area				Urban area			
Sub.	Height	Weight	Chest circumference	Sub.	Height	Weight	Chest circumference
1	163	51	70.00	1	165	65	71.50
2	159	50	71.00	2	167	60	74.50
3	162	51	72.00	3	166	58	79.00
4	161	62	75.00	4	174	52	75.00
5	170	57	81.50	5	178	66	87.00
6	164	55	74.00	6	163	55	69.20
7	168	57	77.00	7	169	56	75.00
8	175	52	88.00	8	164	55	72.00
9	166	54	76.00	9	169	69	80.00
10	157	43	69.00	10	171	56	81.00
11	170	59	73.00				
12	167	57	78.00				

Taking squares on both sides of Eq. (13.11) and making slight changes of the elements, we have

$$t^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' \mathbf{S}_c^{-2} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2). \quad (13.12)$$

If the number of response variables is greater than 1, replace the two sample means in Eq. (13.12) by the mean vectors $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$, and the pooled estimated variance S_c^2 with the variance-covariance matrix \mathbf{S}_c^2 , which is a weighted average of the variance-covariance matrices of the two samples. Then T^2 is generalized as Hotelling's T^2 , that is

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)' \mathbf{S}_c^{-2} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2), \quad (13.13)$$

where

$$\mathbf{S}_c^2 = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1)\mathbf{S}_1^2 + (n_2 - 1)\mathbf{S}_2^2]. \quad (13.14)$$

Under the null hypothesis $H_0 : \mu_1 = \mu_2$, a relationship between T^2 and F value is

$$F = \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 \quad \begin{matrix} v_1 = p \\ v_2 = n_1 + n_2 - p - 1. \end{matrix} \quad (13.15)$$

When the value of $n_1 + n_2$ is large enough, the statistic F follows a χ^2 distribution with p degrees of freedom.

Solution In this example, there are three response variables, height, weight and chest circumference, representing the growth status of young men in this county. We will infer if the mean vectors μ_1 for the rural and μ_2 for the urban are different in this county.

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$

$$n_1 = 12, \quad n_2 = 10, \quad p = 3$$

$$\bar{X}_1 - \bar{X}_2 = \begin{pmatrix} 165.1 \\ 53.9 \\ 75.4 \end{pmatrix} - \begin{pmatrix} 168.6 \\ 59.1 \\ 76.4 \end{pmatrix} = \begin{pmatrix} -3.5 \\ -5.2 \\ -1.0 \end{pmatrix},$$

$$S_1 = \begin{pmatrix} 26.915 & 11.695 & 23.585 \\ 11.695 & 25.201 & 9.259 \\ 23.585 & 9.259 & 28.597 \end{pmatrix},$$

$$S_2 = \begin{pmatrix} 22.142 & 4.842 & 19.842 \\ 4.842 & 30.755 & 13.366 \\ 19.842 & 13.366 & 28.442 \end{pmatrix},$$

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)' S_c^{-1} (\bar{X}_1 - \bar{X}_2) = 10.45,$$

$$F = \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p} T^2 = \frac{18}{60} \times 10.45$$

$$= 3.14, \quad v_1 = 3, v_2 = 18.$$

With $P = 0.051$, H_0 cannot be rejected at the significant level 0.05. We can conclude that the growth status of young individuals in rural and urban areas of this county cannot be considered statistically different.

Consider another situation when we test the differences between the rural and the urban for each response variable separately by repeating the t test, the t values for height, weight and chest circumference will be 1.63, 2.32 and 0.46 with corresponding P values 0.119, 0.031 and 0.653 respectively. The test for height is rejected, but the tests for weight and chest circumference are not. Then we can recognize that the univariate t test can be a supplement to the test of mean vectors rather than a replacement.

13.3 Comparisons among Several Multi-variate Means—Multi-variate Analysis of Variance

Example 13.4 Three groups of data involving two response variables are showed in Table 13.3. Find out if the difference of treatment effects among these three groups is statistically significant.

Table 13.3 The effect scores with two response variables.

Treatment I		Treatment II		Treatment III	
X_i	X_k	X_i	X_k	X_i	X_k
8	2	2	5	3	8
7	4	3	1	2	7
6	3			1	6

Table 13.4 Decomposition of variation for MANOVA.

Source	DF	Sum of squares (Matrix)
Between group	$G - 1$	$H = \sum_{g=1}^G n_g (\bar{X}_g - \bar{X})(\bar{X}_g - \bar{X})'$
Within group	$\sum_{g=1}^G n_g - G$	$E = \sum_{g=1}^G \sum_{j=1}^{n_g} (X_{gj} - \bar{X}_g)(X_{gj} - \bar{X}_g)'$ $= \sum_{g=1}^G (n_g - 1)S_g^2$
Total	$\sum_{g=1}^G n_g - 1$	$H + E$

For a single response, to compare the effect of G (>2) groups, we employed a univariate analysis of variance (ANOVA) introduced in Chap. 12. Similarly, now we would use multi-variate analysis of variance (MANOVA) when there are more than one response variables. The idea of MANOVA is basically the same as that of ANOVA. The total sum of squares of observed effect SS_T is decomposed into two parts, SS_B and SS_W , representing the variation between and within treatment groups. The only difference between these two methods is that SS_T , SS_B , and SS_W in multi-variate analysis are matrices instead of single numbers. In addition, another test statistics Λ^* will be introduced.

13.3.1 Decomposition of variation (matrix)

Table 13.4 shows the decomposition of total variation. The items are similar to ANOVA: n_g is the number of observations in group g . X_{gj} represents the observed vector of the j th subject in group g , \bar{X}_g for the mean vector of group g and \bar{X} for the mean vectors of all observations. H (between

groups) and E (within group) correspond to SS_B and SS_W respectively in ANOVA. $(\cdots)(\cdots)'$ represents the multiplication of row and column vectors, resulting in a matrix of dimension $p \times p$.

For the data in Table 13.3, we have

$$n_1 = 3, n_2 = 2, n_3 = 3; \bar{\mathbf{X}}_1 = \begin{pmatrix} 7 \\ 3 \end{pmatrix},$$

$$\bar{\mathbf{X}}_2 = \begin{pmatrix} 2.5 \\ 3 \end{pmatrix}, \bar{\mathbf{X}}_3 = \begin{pmatrix} 2 \\ 7 \end{pmatrix},$$

$$\bar{\mathbf{X}} = \frac{1}{8} \begin{pmatrix} 8 + 7 + 6 + 2 + 3 + 3 + 2 + 1 \\ 2 + 4 + 3 + 5 + 1 + 8 + 7 + 6 \end{pmatrix} = \begin{pmatrix} 4.0 \\ 4.5 \end{pmatrix},$$

$$\bar{\mathbf{X}}_1 - \bar{\mathbf{X}} = \begin{pmatrix} 3 \\ -1.5 \end{pmatrix}, \bar{\mathbf{X}}_2 - \bar{\mathbf{X}} = \begin{pmatrix} -1.5 \\ -1.5 \end{pmatrix}, \bar{\mathbf{X}}_3 - \bar{\mathbf{X}} = \begin{pmatrix} -2 \\ 2.5 \end{pmatrix},$$

$$\begin{aligned} H &= 3 \begin{pmatrix} 3 \\ -1.5 \end{pmatrix} (3 \quad -1.5) + 2 \begin{pmatrix} -1.5 \\ -1.5 \end{pmatrix} (-1.5 \quad -1.5) \\ &\quad + 3 \begin{pmatrix} -2 \\ 2.5 \end{pmatrix} (-2 \quad 2.5) \\ &= \begin{pmatrix} 43.5 & -24 \\ -24 & 30 \end{pmatrix}. \end{aligned}$$

The covariance matrix of each group is

$$\mathbf{S}_1^2 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}, \mathbf{S}_2^2 = \begin{pmatrix} 0.5 & -2 \\ -2 & 8 \end{pmatrix}, \mathbf{S}_3^2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

$$\mathbf{E} = \sum_{g=1}^3 (n_g - 1) \mathbf{S}_g^2 = 2 \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix} + \begin{pmatrix} 0.5 & -2 \\ -2 & 8 \end{pmatrix}$$

$$+ 2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 4.5 & -1 \\ -1 & 12 \end{pmatrix},$$

$$\mathbf{H} + \mathbf{E} = \begin{pmatrix} 43.5 & -24 \\ -24 & 30 \end{pmatrix} + \begin{pmatrix} 4.5 & -1 \\ -1 & 12 \end{pmatrix} = \begin{pmatrix} 48 & -25 \\ -25 & 42 \end{pmatrix}.$$

Arranging all these results in the format of Table 13.4, we have Table 13.5

Table 13.5 MANOVA of data in Table 13.2.

Source	DF	Sum of squares (matrix)
Between groups	2	$H = \begin{pmatrix} 43.5 & -24 \\ -24 & 30 \end{pmatrix}$
Within group	5	$E = \begin{pmatrix} 4.5 & -1 \\ -1 & 12 \end{pmatrix}$
Total	7	$H + E = \begin{pmatrix} 48 & -25 \\ -25 & 42 \end{pmatrix}$

13.3.2 Wilks' Λ^*

It is a generalized variance ratio and named Wilks' Lambda statistic proposed by Wilks.

$$\Lambda^* = \frac{|E|}{|H + E|},$$

(13.16)

where the numerator and denominator are both determinants. A very small Λ^* implies that the between groups variation H is larger than that of random effect E , indicating that the null hypothesis $H_0 : \mu_1 = \mu_2 = \dots = \mu_G$ is suspectable. The critical point of Λ^* can be obtained by transforming the distribution of Λ^* to a distribution of F following the rules showed in Table 13.6.

We can see from Table 13.6 that when mean vectors of the two groups are to be compared, we can use either Hotelling's T^2 or MANOVA, which is just the same situation with t test and ANOVA in univariate analysis. Although the calculation of MANOVA is considerably cumbersome, statistical packages such as SAS can make it easy. Moreover, the difference between the mean vectors can be analyzed under the multi-variate design. Within the output of SAS procedure, Λ^* can be transformed into F under some commonly appearing conditions. The calculation of Λ^* using the data in Table 13.5 is

$$\Lambda^* = \frac{|E|}{|H + E|} = \frac{\begin{vmatrix} 4.5 & -1 \\ -1 & 12 \end{vmatrix}}{\begin{vmatrix} 48 & -25 \\ -25 & 42 \end{vmatrix}} = 53/1391 = 0.0381.$$

Table 13.6 Distribution of Wilk's Lambda.

No. of variables	No. of groups	Transformation between F and Λ^*	DF
$p = 1$	$G \geq 2$	$F = \left(\frac{\sum n_g - G}{G - 1} \right) \left(\frac{1 - \Lambda^*}{\Lambda^*} \right)$	$v_1 = G - 1,$ $v_2 = \sum n_g - G$
$p = 2$	$G \geq 2$	$F = \left(\frac{\sum n_g - G - 1}{G - 1} \right) \left(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}} \right)$	$v_1 = 2(G - 1),$ $v_2 = 2 \left(\sum n_g - G - 1 \right)$
$p \geq 1$	$G = 2$	$F = \left(\frac{\sum n_g - p - 1}{p - 1} \right) \left(\frac{1 - \Lambda^*}{\Lambda^*} \right)$	$v_1 = p,$ $v_2 = \sum n_g - p - 1$
$p \geq 1$	$G = 3$	$F = \left(\frac{\sum n_g - G}{p} \right) \left(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}} \right)$	$v_1 = 2p,$ $v_2 = 2(n_g - p - 2)$

Referring to Table 13.5, $p = 2$, $G = 3$, and get F value

$$\begin{aligned}
 F &= \left(\frac{\sum n_g - G - 1}{G - 1} \right) \left(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}} \right) \\
 &= \left(\frac{8 - 3 - 1}{3 - 1} \right) \left(\frac{1 - \sqrt{0.0381}}{\sqrt{0.0381}} \right) = 8.2463,
 \end{aligned}$$

$$v_1 = 2(G - 1) = 4, \quad v_2 = 2 \left(\sum n_g - G - 1 \right) = 8.$$

Then, we have the corresponding P value from the F table, $P = 0.0061$. Since $P < 0.01$, the null hypothesis of $H_0 : \mu_1 = \mu_2 = \mu_3$ is rejected, and the effectiveness scores of these three therapies are considered significantly different.

If x_1 and x_2 were analyzed separately with ANOVA, $F = 24.17$, $P = 0.0027$ for x_1 , and $F = 6.25$, $P = 0.0436$ for x_2 would be obtained. Thus,

we cannot reach an overall conclusion about the difference among the three treatments.

13.4 Computerized Experiments

Experiment 13.1 Hotelling's T^2 test for $H_0 : \boldsymbol{\mu} = \mathbf{0}$ Program 13.1 is applied to examine the effect of lowering blood lipid of the drug in Example 13.2. The hypothesis to be tested is $H_0 : \boldsymbol{\mu} = (0 \ 0)'$. Hotelling's T^2 is generated by the procedure GLM in SAS. The value of F defined in Eq. (13.10) is obtained by transforming the statistics listed below:

- | | |
|----------------------------|--|
| (1) Wilks Λ^* | Wilks' Lambda = $ E / H + E $ |
| (2) Pillai trace | Pillai's trace = $\text{trace}(H(H + E)^{-1})$ |
| (3) Hotelling-Lawley trace | Hotelling-Lawley trace = $\text{trace}(E^{-1}H)$ |
| (4) Greatest root of R | Roy's greatest root of $E^{-1}H$. |

Note that the values of F derived from each of the four transformations above may not be identical.

$C = 1$ in line 03 implies that there is only one group. Data in line 05 is the observed data in Table 13.1. Lines 07–11 are the statement GLM to perform the analysis. In line 09, x_1 and x_2 are placed on the left-hand side of equitation to indicate that it is a multi-variate model; the option NOUNI instructs the program not to display univariate ANOVA results on the response variables x_1 and x_2 separately. Line 10 instructs SAS to test whether the population mean vector is equal to zero. LSMEANS in line 11 asks the computer to display the mean, standard error and P value for testing whether the population mean is zero. Lines 12 and 13 perform the procedure

Program 13.1 Test for a zero mean vector.

Line	Program	Line	Program
01	DATA HOTE;	08	CLASS C;
02	INPUT X1 X2 @@;	09	MODEL X1 X2=C/NOUNI;
03	C=1;	10	MANOVA H= INTERCEPT;
04	CARDS;	11	LSMEANS C/STDERR PDIFF;
05	16 -4 21 46 57 -40 -20 107 17 86	12	PROC CORR COV OUTP=A;
06	;	13	VAR X1 X2;
07	PROC GLM;	14	PROC PRINT; RUN;

Table 13.7 Result of MANOVA.

Statistic	Value	<i>F</i>	Num <i>DF</i>	Den <i>DF</i>	Pr> <i>F</i>
Wilks' Lambda	0.131278	9.9261	2	3	0.0476
Pillai's Trace	0.868722	9.9261	2	3	0.0476
Hotelling-lawley Trace	6.617427	9.9261	2	3	0.0476
Roy's Greatest Root	6.617427	9.9261	2	3	0.0476

Table 13.8 Result of *t* tests of univariate x_1 and x_2 .

X_1 LSMEAN	Std Err LSMEAN	Pr > <i>t</i> H0:LSMEAN=0	X_2 LSMEAN	Std Err LSMEAN	Pr > <i>t</i> H0:LSMEAN=0
18.2	12.2040977	0.2101	39	27.3605555	0.2272

Table 13.9 Basic multi-variate statistics.

OBS	_TYPE_	_NAME_	X_1	X_2
1	COV	X_1	744.70	-1401.25
2	COV	X_2	-1401.25	3743.00
3	MEAN		18.20	39.00
4	STD		27.29	61.18
5	N		5	5
6	CORR	X_1	1	-0.84
7	CORR	X_2	-0.84	1

CORR to produce multi-variate statistics such as mean vectors, covariance matrix and correlation matrix of X_i and X_k introduced in Sec. 13.1. The main results are showed in Tables 13.7, 13.8 and 13.9.

Experiment 13.2 Testing the difference of mean vectors of multi-group observations Program 13.2 is used to infer the statistical difference of mean vectors of $G (\geq 2)$ groups. The value of F is defined in Eq. (13.15) and Table 13.6 is also derived from transformations mentioned in Experiment 13.1. AREA, H, W and B in line 02 represent the grouping variable area as well as the variables to be investigated, namely, body height, weight and chest circumference. Lines 04–14 show the observed data. Lines 16–19 perform the test of mean vectors between the rural and the urban. Line

Program 13.2 Test for mean vectors of multi-groups.

Line	Program	Line	Program
01	DATA GROWTH;	13	T 169 56 75.0 T 164 55 72.0
02	INPUT AREA \$ H W B @@;	14	T 169 69 80.0 T 171 56 81.0
03	CARDS;	15	;
04	R 163 51 70.0 R 159 50 71.0	16	PROC GLM; CLASS AREA;
05	R 162 51 72.0 R 161 62 75.0	17	MODEL H W B=AREA/NOUNI;
06	R 170 57 81.5 R 164 55 74.0	18	MANOVA H=AREA/PRINTE PRINTH;
07	R 168 57 77.0 R 175 52 88.0	19	LSMEANS AREA/STDERR PDIFF ;
08	R 166 54 76.0 R 157 43 69.0	20	PROC SORT; BY AREA;
09	R 170 59 73.0 R 167 57 78.0	21	PROC CORR COV OUTP=A;
10	T 165 65 71.5 T 167 60 74.5	22	VAR H W B; BY AREA;
11	T 166 58 79.0 T 174 52 75.0	23	PROC PRINT; RUN;
12	T 178 66 87.0 T 163 55 69.2		

18 represents the test of mean vectors between subjects of individuals in different areas by MANOVA, with the options of PRINTE and PRINTH requiring the display of vector H and E respectively. When there are more than three treatment groups, the statement LSMEANS in line 19 provides the results of comparisons among the means of all treatment groups and zero,

Table 13.10 Results of MANOVA (extract).

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.66146388	3.07	3	18	0.0542

Table 13.11 Results of uni-variate analysis and multiple comparison (from statement LSMEANS).

VAR	AREA	LSMEAN	Std Err LSMEAN	Pr > T H_0 :LSMEAN=0	Pr > T H_0 :LSMEAN1 =LSMEAN2
H	R	165.166667	1.431879	<0.0001	0.1216
	T	168.600000	1.568545	<0.0001	
W	R	54.000000	1.5297059	<0.0001	0.0329
	T	59.200000	1.6757088	<0.0001	
H	R	75.375000	1.5418313	<0.0001	0.6526
	T	76.420000	1.6889916	<0.0001	

Table 13.12 General statistics (from procedure CORR).

AREA	_TYPE_	_NAME_	H	W	B
R	COV	H	26.69697	11.72727	23.34091
R	COV	W	11.72727	25.09091	9.409091
R	COV	B	23.34091	9.409091	28.59659
R	MEAN		165.1667	54	75.375
R	STD		5.166911	5.009083	5.347578
R	N		12	12	12
R	CORR	H	1	0.453114	0.844753
R	CORR	W	0.453114	1	0.351263
R	CORR	B	0.844753	0.351263	1
T	COV	H	22.04444	4.866667	19.83111
T	COV	W	4.866667	31.73333	13.65111
T	COV	B	19.83111	13.65111	28.44178
T	MEAN		168.6	59.2	76.42
T	STD		4.695151	5.633235	5.333083
T	N		10	10	10
T	CORR	H	1	0.184003	0.791989
T	CORR	W	0.184003	1	0.454393
T	CORR	B	0.791989	0.454393	1

and all pairwise comparisons with exact P values displayed. There are only two groups in this example so that this statement will give the results of two t tests, showed in Tables 13.10 and 13.11. In order to acquire the corresponding mean vector, covariance and correlation matrixes of each group, sort the observations by SORT statement in advance and then obtain the general multi-variate statistics illustrated in Sec. 13.1 by applying procedure CORR in lines 21–23. The mean vectors, covariance matrix and correlation matrix of H, W, B of each area group are showed in Table 13.12.

13.5 Practice and Experiments

1. Analyze the laboratory results on ferrohemoglobin and RBC of patients with anemia presented in Table 13.13 (from Shi Bingzhang, Yang Qi, Medical Multi-variate Analysis, People's Medical Publishing House, 1990).

- (1) Compute multi-variate statistics for each patient group A, B and C, and make brief descriptions about the differences among the groups and the relationships among response variables.

Table 13.13 Ferrohemoglobin and RBC of patients with anemia.

Group A ($n_1 = 12$)		Group B ($n_2 = 10$)		Group C ($n_3 = 8$)	
ferrohemoglobin (g/L)	RBC (1012/L)	ferrohemoglobin (g/L)	RBC (1012/L)	ferrohemoglobin (g/L)	RBC (1012/L)
39	2.1	48	2.7	44	2.5
42	1.9	47	1.8	37	3.0
37	2.4	54	2.3	29	2.4
40	1.7	45	2.4	45	3.3
44	2.2	46	2.7	33	2.3
52	2.3	44	2.2	45	1.9
27	1.6	59	2.9	38	2.7
24	2.6	55	2.2	37	3.1
36	2.4	43	2.9		
55	1.8	51	3.1		
29	2.0				
33	3.0				

(2) Infer the difference among patient groups with MANOVA.

2. Test the difference among the treatment effects of the three groups of patients in Example 13.4 by modifying procedure 13.2. If the statement in line 19 of program 13.2 is replaced by `MANOVA H = Δ/PRINTE PRINTH` (note: Δ is the name of the variable defined by your self), SAS will provide the within group (error) sum of squares matrix E and between groups (treatment effect) sum of squares matrix H . Compare this output with that in Table 13.5, think and discuss the similarities and differences between ANOVA and MANOVA.

3. Analyze the laboratory results on the improvement of immune globulin after treatment of thymosin presented in Table 13.14 (from Shi Bingzhang, Yang Qi, and Medical Multi-variate Analysis. The People's Health Press, 1990).

- (1) Use univariate procedures to test the improvement after treatment on IgG, IgA and IgM respectively.
- (2) Are there any disadvantages or weaknesses if this data are analyzed by univariate statistics?
- (3) Consider the three measurements of each patient as a vector, calculate the mean vector and covariance matrix.

Table 13.14 Improvement of immune globulin after treatment of thymosin.

No. of patient	IgG (g/L)	IgA (g/L)	IgM (g/L)
1	-1.56	-500	-490
2	-1.76	-50	-140
3	-0.63	-120	-210
4	-1.28	-700	90
5	0.07	150	-180
6	-1.42	-620	190
7	-1.04	740	-240
8	-1.95	110	-40
9	-4.20	-540	160
10	-2.36	-600	-380
11	-2.14	-880	-220
12	-1.39	110	-220
13	-0.71	90	110
14	-1.56	-310	-40
15	-0.49	-50	-200

(4) Infer the effect of thymosin in improving immune globulin with Hotelling T^2 test.

(1st edn. Yongyong Xu, Jiqian Fang, Danhong Liu; 2nd edn. Yongyong Xu, Danhong Liu)

Chapter 14

Discriminant Analysis

Discriminant analysis is a technique for classifying subjects into different groups according to the measured covariates. In clinical research, doctors make differential diagnosis based on the symptoms, laboratory results, pathological tests, and imaging reports. Sometimes, doctors want to further classify a patient into different subtypes or disease stages. All these problems can be resolved using discriminant analysis. In addition, a discriminant analysis can estimate relative contributions of covariates.

14.1 Basic Ideas of Discriminant Analysis

In the next section, we will use an example to illustrate the basic idea of discriminant analysis. We will use microspectrofluorophotometry to analyze cells from subjects with cancerous tumor and control subjects with no cancerous tumor. For each subject, we have three measures: X_1 the score of triploids, X_2 the score of octalpoids, and X_3 the score of aneuploids. Based on observed data (or a training sample), a discriminant function $Y = X_1 + 10X_2 + 10X_3$ is developed. For a subject with values of (X_1, X_2, X_3) , we can substitute these values to derive the corresponding values of discriminant function Y . If $Y > 100$, the subject is classified as cancer subject; otherwise, for $Y < 100$, the subject is classified as no cancerous tumor. This discriminant function and the corresponding decision rule need to be further validated based on another real data. When the accuracy of the decision rule meets the clinical requirements after validation, it can be used in clinical practice.

Table 14.1 Data structure of training data.

Subject Id	Explanatory variables						Classification variables (Y)
	X_1	X_2	...	X_j	...	X_p	
1	X_{11}	X_{12}	...	X_{1j}	...	X_{1p}	Y_1
2	X_{21}	X_{22}	...	X_{2j}	...	X_{2p}	Y_2
...
i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{ip}	Y_i
...
n	X_{n1}	X_{n2}	...	X_{nj}	...	X_{np}	Y_n

In discriminant analysis a discriminant function is used to discriminate subjects. A training sample is used to establish such a discriminant function. The goal of the discriminant analysis is to identify and evaluate performances of different and possible discriminant functions in the training sample. Thus, the quality and the size of the training sample are critical to establish useful discriminant rules. The classification of each subject in the training sample should be known according to the “gold standard.” Multiple explanatory variables, denoted as variables X_1, X_2, \dots, X_p , that are related by chance into different classes are collected. These variables have to be measured accurately and the number of subjects should be sufficiently large.

Table 14.1 gives the data structure of a training sample with Y indicating the classes from $1, 2, \dots, g$; and X_1, X_2, \dots, X_p , as explanatory variables. We use the subscript i to indicate data from the i th subjects.

Several discriminant analysis methods are available. The common ones are the following:

- (1) Maximum likelihood method: This method evaluates the likelihood of being in each group among possible choices. It is suitable for parametric and semi-parametric models.
- (2) Fisher’s discriminant method: This method is commonly used to discriminate two groups. In the example of Table 14.1 Fisher’s discriminant method is used.
- (3) Bayesian discriminant method: This method required multivariate normal distributions of explanatory variables and is used to discriminate two or more groups.
- (4) Logistic discriminant method: This method is commonly used to discriminate two groups. It does not require normal distribution

assumptions. Instead of modeling the chance of falling into one of the two groups, this method can also be used for ordinal or binary explanatory variables.

This chapter focuses on the introduction of Fisher's and Bayes discriminant methods, two most widely used methods.

14.2 Fisher's Discriminant Analysis

Let us assume that there are two explanatory variables: X_1 (temperature) and X_2 (protoheme counts) that are used to discriminate two diseases A and B . In Fig. 14.1(a), each point represents a subject. If we compare the temperature between two disease types, the difference is minimal. Similarly, minimal difference is observed in protoheme counts. However, combining two variables makes it relatively easy to separate the two groups. If we can project our observations to a direction that the projected values can easily separate the two groups, we can use the projected values to make decisions. For a new subject, we can based on the projection of the temperature and protoheme counts to determine the disease status of the subject. What is the best projection direction? If we use the principle of the analysis of variance, we want to maximize the ratio of the between-class variation to the within-class variation. This is the basic idea of Fisher's discriminant function.

14.2.1 Discrimination of two groups

Suppose that we want to separate populations π_1 and π_2 . We assume the covariance matrices of explanatory variables are the same for the

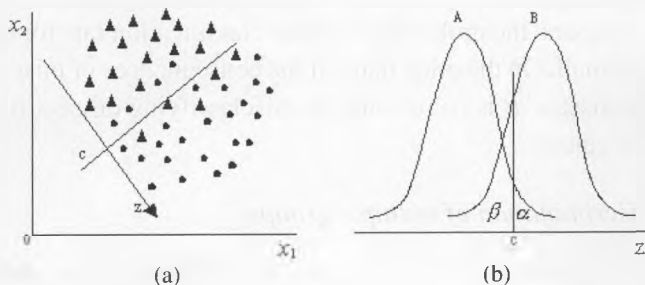


Fig. 14.1 A sketch map for Fisher discriminant analysis (a) two-dimensional scatter plot: less overlap along the direction, (b) the distribution of Z and the cutoff point.

populations π_1 and π_2 . We randomly select individuals, several from each population, as the training sample. Fisher (1936) proposed to use coefficients a_1, a_2, \dots, a_p to construct a score Z :

$$Z = a_1 X_1 + a_2 X_2 + \dots + a_p X_p \quad (14.1)$$

which maximizes the distance between mean scores of the two groups,

$$D^2 = \frac{(\bar{Z}_1 - \bar{Z}_2)^2}{S_Z^2} \quad (14.2)$$

Here, \bar{Z}_1 and \bar{Z}_2 are sample means of Z from two populations and S_Z^2 is the pooled estimation of sample variances. D^2 is also called Mahalanobis distance.

Once we find coefficients a_1, a_2, \dots, a_p that maximize D^2 , the corresponding function Z in (14.1) is called discriminant function. The coefficients are called discriminant coefficients. The discriminant function is the weighted average of all explanatory variables with the weight of discriminant coefficients, which maximally separate the two groups. Thus, multiple explanatory variables are reduced into a univariate function Z .

Classifying a subject into one of the two groups from a univariate function Z is not difficult. We only need to determine a threshold value C , such that all subjects with $Z \leq C$ are classified as one group and subjects with $Z > C$ are in another group.

To determine the threshold C , we can use the histograms of Z from two populations (Fig. 14.1(b)). Depending on clinical problems, we can decide the threshold with corresponding false classification probabilities, α and β , in Fig. 14.1(b). For example, if we take benign tumors as A and malignant tumors as B , the probability of false classification for disease A , α could be relatively larger and the probability of false classification rate for disease B , β , should be small. On the other hand, if the consequences of misclassifying disease A to disease B is comparable to misclassifying disease B into A , α and β can be equal.

14.2.2 Discrimination of multiple groups

When we have more than two groups, we are still looking for a linear combination of explanatory variables such that overlaps of projected values from these groups are minimized. The overlap is measured by the ratio of

between group variance over the within group variance. When the number of explanatory variables is more than the number of groups, i.e. $P \geq g$, there are $g - 1$ discriminant functions. Similar to two groups situation, these discriminant functions only provide discriminant scores, not the discriminant rules. We have to determine the classification rules according to clinical properties of these groups. As it is easier for multiple groups to overlap with each other, discriminant analysis of multiple groups is usually less efficient than the classification of the two groups.

14.3 Bayesian Discriminant Analysis

14.3.1 Bayesian criterion

Assume that there are g populations $\pi_1, \pi_2, \dots, \pi_g$ and the explanatory variables in each population, X_1, X_2, \dots, X_p follow multivariate normal distributions. For each subject with observations of p variables, we want to determine which of g populations the subject belongs to.

When we make a classification decision, we inevitably make errors of misclassification. Let the probability of misclassifying a subject in class i into class j as $P(i|j)$ and the loss due to the misclassification as $C(j|i)$. The Bayesian criterion is to minimize the expected misclassification loss. The discriminant rule based on this criterion is the Bayesian discriminant analysis.

14.3.2 Classification function

The Bayesian classification function is as follows:

$$\begin{aligned} Y_1 &= C_{10} + C_{11}X_1 + C_{12}X_2 + \dots + C_{1p}X_p, \\ Y_2 &= C_{20} + C_{21}X_1 + C_{22}X_2 + \dots + C_{2p}X_p, \\ &\vdots \\ Y_g &= C_{g0} + C_{g1}X_1 + C_{g2}X_2 + \dots + C_{gp}X_p. \end{aligned} \tag{14.3}$$

There exist g linear equations with each linear equation for one of the populations. Here, $C_{j0}, C_{j1}, \dots, C_{jp}$ ($j = 1, 2, \dots, g$) are the parameters to be estimated and Y_j is positively related to the probability of being in the j th population. SAS procedure DISCRIM can estimate these coefficients. After classification functions are established, the discriminant rules

are developed by substituting the observed values of explanatory variables into Eq. (14.3). A subject is classified into population f that gives the largest Y value.

14.3.3 Prior probability

A prior probability of the subpopulation Y_i , $q(Y_i)$, is the probability of randomly selecting a sample of j th subpopulation from the whole population. For example, of all patients of appendicitis, 50% are contiguity, 30% are abscess, 10% are gangrene, and 10% are peritonitis. Therefore, if we randomly select a patient with appendicitis, the probability of a patient from these four types are about 0.5, 0.3, 0.1, and 0.1, respectively. Classification function (14.3) did not consider the prior probability. When these prior probabilities are considered, the classification functions become the following:

$$\begin{aligned} Y_1 &= C_{10} + C_{11}X_1 + C_{12}X_2 + \cdots + C_{1P}X_P + \ln(q(Y_1)), \\ Y_2 &= C_{20} + C_{21}X_1 + C_{22}X_2 + \cdots + C_{2P}X_P + \ln(q(Y_2)), \\ &\vdots \\ Y_g &= C_{g0} + C_{g1}X_1 + C_{g2}X_2 + \cdots + C_{gP}X_P + \ln(q(Y_g)). \end{aligned} \quad (14.4)$$

The only differences between (14.3) and (14.4) are the additions of $\ln(q(Y_j))$.

By including the prior probabilities, the sensitivity of discriminant analysis can be improved. However, the prior probability is often unknown and difficult to estimate. If the training samples are randomly selected from the whole population, instead of stratified samples from each subpopulation, we can use the observed frequency of the subpopulation of Y_j , $Q(Y_i)$, to estimate the prior probability of $q(Y_j)$. When there is no information about prior probability and the frequencies of subpopulations $Q(Y_i)$ cannot be used to estimate $q(Y_j)$, a conservative approach is to set equal prior, i.e. to set $q(Y_j) = 1/g$.

14.3.4 Posterior probability

A posterior probability of a sample is the probability of a sample belonging to subpopulation Y_j after observing values S_i of the explanatory variables X_i . It is often denoted as $P(Y_i|S_1, S_2, \dots, S_p)$.

Once we observed the explanatory variables X_1, X_2, \dots, X_P of a sample, we can calculate the posterior probability for subpopulation Y_j . Thus, we have a quantitative indicator about how likely the sample belongs to a specific subpopulation. By substituting observed values S_1, S_2, \dots, S_p into classification function (14.4), we can calculate Y_1, Y_2, \dots, Y_g . The posterior probabilities are calculated using the following formulas:

$$\begin{aligned} P(1st|S_1, S_2, \dots, S_p) &= \frac{\exp(Y_1)}{\sum_{j=1}^g \exp(Y_j)}, \\ P(2nd|S_1, S_2, \dots, S_p) &= \frac{\exp(Y_2)}{\sum_{j=1}^g \exp(Y_j)}, \\ &\vdots \\ P(gth|S_1, S_2, \dots, S_p) &= \frac{\exp(Y_g)}{\sum_{j=1}^g \exp(Y_j)}. \end{aligned} \quad (14.5)$$

When Y_j is too small or too large, we may experience numerical floating errors. To avoid this problem, we can add or subtract a constant from all Y_j 's. For example, by subtracting $Y^* = \max(Y_1, Y_2, \dots, Y_g)$ and recalculating the posterior probability, the equations in (14.5) become

$$P(jth|S_1, S_2, \dots, S_p) = \frac{\exp(Y_j - Y^*)}{\sum_{j=1}^g \exp(Y_j - Y^*)} \quad j = 1, 2, \dots, g. \quad (14.6)$$

Classification of a sample based only on the values of posterior probabilities is insufficient. For example, when the posterior probabilities of a sample from three subpopulations are 0.95, 0.03, and 0.02 respectively, the confidence of classifying the sample to subpopulation 1 is reasonably higher. However, when the posterior probabilities are 0.4, 0.3, and 0.3, respectively, the confidence of classifying the sample to subpopulation 1 is low. In clinical diagnosis for samples with questionable confidence, we can assign the samples as to be determined. The PROC DISCRIM procedure in SAS can define a threshold of posterior probabilities. When the highest posterior probability is above the threshold, we can make classification decisions. Otherwise, the sample will be classified as others for further determination.

Example 14.1 To study the retinopathy severity (mild, moderate, and severe) among diabetic patients, researchers collected patients age (age),

disease duration in years (time), glucose level, vision, the peak time (at) and the amplitude (av) of *a*-wave, the peak time (bt) and amplitude (bv) of *b*-wave, and the peak time (qpt) and amplitude (qp) of *qp*-wave from an electro-retinography. The goal of the study was to use these variables to establish a three-level severity index of retinopathy. A total of 131 diabetic patients were examined. The inclusion criteria include no previous diagnosed eye diseases other than retinopathy due to diabetes. The data was given in Appendix III. The severity of retinopathy was named as "group". We used this group of 131 patients as training sample to discriminate severity groups according to age, vision, at, bt, and qp. The question is for Mr. Smith, who is 38 years old with vision of 1.0, at = 14.25, bv = 383.39, and qp = 43.18, which is his most likely severity group?

Solution Suppose the 131 patients were random samples from the target population: diabetic patients with retinopathy. As random samples can be used to approximate the population, we can estimate the prior probability. Using PROC DISCRIM procedure of SAS package, we have the following classification function:

$$Y_1 = -181.447 + 0.473(\text{age}) + 60.369(\text{vision})$$

$$+ 17.708(\text{at}) + 0.048(\text{bv}) + 0.364(\text{qp}),$$

$$Y_2 = -165.830 + 0.472(\text{age}) + 49.782(\text{vision})$$

$$+ 17.658(\text{at}) + 0.034(\text{bv}) + 0.325(\text{qp}),$$

$$Y_3 = -189.228 + 0.178(\text{age}) + 43.974(\text{vision})$$

$$+ 20.447(\text{at}) + 0.040(\text{bv}) + 0.265(\text{qp}).$$

According to the observed measures of Mr. Smith, the classification functions are $Y_1 = 183.36$, $Y_2 = 180.58$, and $Y_3 = 179.66$. Y_1 is the largest among the three. Thus, we will place Mr. Smith into the mild severity group.

Although the values of Y_1 , Y_2 and Y_3 are quite similar, a decision based only on the values of classification functions ignores the relative differences between these values. Estimation of the posterior probability will provide a better picture of his classification. In order to avoid numerical problem in computation, we let $Y^* = 180$. The posterior probability can be calculated

as follows:

$$\begin{aligned}
 P(1st|X_1, X_2, \dots, X_5) &= \frac{\exp(183.36 - 180)}{\exp(183.36 - 180) + \exp(180.58 - 180) + \exp(179.66 - 180)} \\
 &= 0.9202.
 \end{aligned}$$

Similarly,

$$P(2nd|X_1, X_2, \dots, X_5) = 0.0571,$$

$$P(3rd|X_1, X_2, \dots, X_5) = 0.0227.$$

Thus, we can confidently assign Mr. Smith to the mild diseased group.

14.4 Stepwise Discriminant Function

The multiple regression analysis already taught us that the increased number of independent variables in a regression equation does not necessarily correspond to a better model. Insignificant variables will not help in explaining the dependent variable but rather introducing more noise to the model. Similarly, in a discriminant analysis, it is not always true that the more variables there are, the better the discriminant function is. On the contrary, we want to avoid explanatory variables that make minimum contributions and keep only those explanatory variables that can address the most differences between groups. Similar to the stepwise regression, stepwise discriminant analysis is likely to select significant explanatory variables into the model and keep insignificant variables away from the final discriminant model. The level of importance of an explanatory variable can be tested using an F -test. The null hypothesis of this test is that the contribution of this variable to explained variations between groups is zero. A small P -value will reject the null hypothesis and suggest that the variable contributes significantly to account for the differences between groups. In Example 14.1, we have ten variables. Their F -statistics and P -value of the F -test is given in Table 14.2.

If we use the significance level of 0.05, six of the ten variables have P -values above this. Therefore, we should use stepwise procedure to determine which variables are necessary for discriminant analysis.

Table 14.2 Contribution of ten variables in the discriminant analysis.

Variables	<i>F</i> -statistics	<i>P</i> -values
Age	25.338	0.0001
Duration of diseases (year)	1.211	0.3016
Glucose	1.255	0.2889
Vision	45.956	0.0001
at	20.310	0.0001
av	0.219	0.8037
bt	0.950	0.3898
bv	6.012	0.0033
qpt	0.971	0.3818
qpv	1.989	0.1414

PROC STEPDISC in SAS can perform stepwise selection of explanatory variables. Its step and procedure are similar to stepwise regression. First, we need to select the significance level to select and remove variables (i.e. type I error level), denoted as P_1 for the selection and P_2 for the removal. P_1 and P_2 can be equal or different, such as 0.05, 0.1, or 0.15. In a case that P_1 and P_2 are different, P_1 is usually smaller than P_2 . The smaller P_1 is, the fewer variables are included into the discriminant function. Further, the selection of inclusion variables is also step by step. In each step, we always select the variable of the smallest P -value that is smaller than P_1 . When a new variable is included, we will revisit the discriminant function to examine the significance of all variables already in the model. If one or more variables have P -values greater than P_2 , these variables will be removed from the discriminant function. When there are no more variables to enter and no more variables to remove, the stepwise procedure is complete and the final stepwise function is derived.

Example 14.2 Using training samples in Example 14.1 to run the PROC STEPDISC of SAS, we have five variables retained in discriminant function using both P_1 and P_2 as 0.05. The results are summarized in Table 14.3.

From Table 14.3, the five variables: age, vision, at, bv, and qpv were significant variables for discriminant purpose. After the inclusion of age, the other five variables had no significant contributions and therefore, were excluded from the discriminant function. Finally, we can use SAS procedure

Table 14.3 Results of stepwise discriminant function.

Included in discriminant function			Excluded in discriminant function		
Variables	F-value	P-value	Variables	F-value	P-value
Age	28.818	0.0001	Duration	0.891	0.4127
Vision	46.491	0.0001	Glucose	0.793	0.4548
at	24.964	0.0001	av	0.397	0.6730
bv	9.387	0.0002	bt	0.421	0.6572
qpv	3.829	0.0243	qvt	1.016	0.3649

DISCRIM to recalculate classification functions using age, at, bv, and qpv (Example 14.1).

14.5 Decision Tree

Decision tree-based method is a kind of nonlinear discriminant analysis. With a set of partition rules, the probability distribution of all possible outcomes is expressed by a decision tree, in order to predict or classify the individuals' categories as accurate as possible. The main applications of decision tree in the medical field throughout medical diagnosis to resource allocation and much wider with the rapid development of information technology.

14.5.1 Basic idea of decision tree approach

Before introducing the methods of decision tree, let us have the definition of a "Tree" first. A tree is a nonempty set of nodes, which consists of three kinds: root nodes, internal nodes and leaf nodes (also called terminal nodes). Root node is the top node of a tree, and each internal node represents an attribute, while each branch corresponds to one output of the attribute and each leaf node represents one category.

Numbers of layers for a decision tree are diverse under different circumstances. Generally, one decision tree has only one root node, which can be considered as an internal node or a parent node. One parent node will be divided into two daughter nodes, which are called left daughter node and right daughter node, respectively. Figure 14.2 is an example of tree including four layers: three internal nodes (including the root node) with label 1,

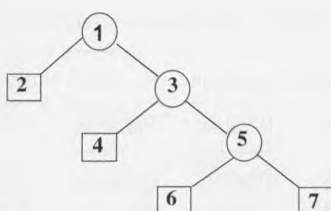


Fig. 14.2 An example of tree diagram.

3, 5 (circled) and four terminal nodes with label 2, 4, 6, 7 (square). Each internal node is connected with its own daughter nodes, and so terminal nodes have no offspring (means daughter node). Besides, a tree diagram might be asymmetric because a daughter node may be internal node or terminal node. See Fig. 14.2, node 2 is a terminal node while node 3 is internal nodes, both of them are the daughter nodes of node 1. It depends on the splitting rule to divide parent node into two daughter nodes accurately.

Fundamental principle of decision tree method: Start from a root note to construct each layer of a tree by a top-down splitting rules. Firstly, the observed sample (training sample) will be divided into several disjoint subset based on a cutoff of one attribute which has highest information gain. All subjects in each subset have the similar attribute values in the selected attribute, and each subset represents one daughter node for the tree. Secondly, select new attribute to divide each daughter node into secondary daughter nodes, and repeat this process for each subtree until the specified criterion is fulfilled that each terminal node represents one category. The path from the node of the attribute to each daughter node corresponds to the relationship between each testing node and each specific category.

There are many kinds of classification rule for decision tree method. Based on the functions, decision tree can be divided into survival tree, classification tree and regression tree; while based on the numbers of categories, there are binomial classification tree and multinomial classification tree. Due to the characteristics of readability and popularity for binomial classification tree, and one can obtain multinomial classification tree by repeating binomial classification tree. This chapter will mainly focus on binomial classification tree, non-terminal nodes for the binomial classification are only allowed to have two daughter nodes.

14.5.2 Classification and regression tree (CRT)

Although the process of constructing a classification tree is complex, there are generally three major issues needed to solve: 1) What are the nodes? That means to determine the root node, internal nodes and terminal nodes of a tree. 2) How to divide the nodes? That is how to use the training sample to construct a tree from the root node. 3) How to stop a tree growing, which means how to select a best tree based on a training sample.

14.5.2.1 The growth of the tree (Splitting Rule)

A tree starts growing from divided a root node into two daughter nodes, and then the daughter nodes are divided into sub-daughter nodes using the same rule, and iterate this process to grow the tree. The variation of variables used to split sample in each internal nodes will be smaller than those corresponding parent nodes. The independent variables have different splitting methods when there are more than one variable, so evaluation criterion are essential to judge the purity of the node and then make best decision. For ordinal variables, there are $(k - 1)$ methods to divide the variable, where k denotes the number of the possible values. For binomial data, there is only one decision method. But for categorical data with k levels, $2k - 1 - 1$ methods are available. For example, seven decision methods are available for ABO blood types. The results are shown in Table 14.4.

The decision tree is called classification tree when the outcome variables are categorical variables. Three splitting rules can be applied, entropy method, Gini index method and Pearson chi-square test. The following

Table 14.4 All possible decision methods for blood type.

Left daughter nodes	Right daughter nodes
A	B, AB, O
B	A, AB, O
AB	A, B, O
A, B	AB, O
A, AB	B, O
B, AB	A, O
A, B, AB	O

Table 14.5 Cross classification table of the nodes and dependent variable.

Daughter node	Cutoff value	NO	YES	All
Left daughter node (t_L)	$x_j < c$	n_{11}	n_{12}	$n_{1\bullet}$
Right daughter node (t_R)	$x_j \geq c$	n_{21}	n_{22}	$n_{2\bullet}$
Parent node (t)	—	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

example is applied to explain algorithm of these three approaches. Let c denotes the cutoff point of a continuous variable x , and dependent variable y is a binomial variable (0 = NO, 1 = YES). The results are shown in Table 14.5.

a) Entropy reduction

The entropy impurity for left daughter nodes is defined as

$$E(t_L) = -\frac{n_{11}}{n_{1\bullet}} \log \left(\frac{n_{11}}{n_{1\bullet}} \right) - \frac{n_{12}}{n_{1\bullet}} \log \left(\frac{n_{12}}{n_{1\bullet}} \right). \quad (14.7)$$

While the entropy impurity for right daughter nodes and parent nodes are calculated by

$$\begin{aligned} E(t_R) &= -\frac{n_{21}}{n_{2\bullet}} \log \left(\frac{n_{21}}{n_{2\bullet}} \right) - \frac{n_{22}}{n_{2\bullet}} \log \left(\frac{n_{22}}{n_{2\bullet}} \right), \\ E(t) &= -\frac{n_{\bullet 1}}{n_{\bullet\bullet}} \log \left(\frac{n_{\bullet 1}}{n_{\bullet\bullet}} \right) - \frac{n_{\bullet 2}}{n_{\bullet\bullet}} \log \left(\frac{n_{\bullet 2}}{n_{\bullet\bullet}} \right), \end{aligned} \quad (14.8)$$

then we can get the reduction of entropy by Eq. (14.9).

$$\Delta E(t) = E(t) - P(t_L)E(t_L) - P(t_R)E(t_R), \quad (14.9)$$

where

$$P(t_L) = \frac{n_{1\bullet}}{n_{\bullet\bullet}}, \quad P(t_R) = \frac{n_{2\bullet}}{n_{\bullet\bullet}}$$

represent the probability of t_L and t_R , respectively. The entropy reduction $\Delta E(t)$ is an index of goodness of split (also called Information Gain), which represents that the degree to which impurity is reduced due to dividing the parent node into two daughter nodes. Usually we choose the cutoff point of the variable with the largest entropy reduction. If the dependent variable has multiple levels, add corresponding category to the equation. For instance, if

there are k levels for independent variable, and the ratios for each category is $p_i (i = 1, 2, \dots, k)$, then

$$E(t_L) = - \sum_i p_i \ln p_i. \quad (14.10)$$

b) Gini Impurity

The value of Gini index for a completely pure node is 0, and Gini index will approach 1 with the growing of the internal classification of a node. The larger reduction of Gini index, the better the decision is.

The Gini index for left daughter node is

$$G(t_L) = 1 - \left(\frac{n_{11}}{n_{1\bullet}} \right)^2 - \left(\frac{n_{12}}{n_{1\bullet}} \right)^2 \quad (14.11)$$

and for right daughter node and parent node is

$$\begin{aligned} G(t_R) &= 1 - \left(\frac{n_{21}}{n_{2\bullet}} \right)^2 - \left(\frac{n_{22}}{n_{2\bullet}} \right)^2, \\ G(t) &= 1 - \left(\frac{n_{\bullet 1}}{n_{\bullet\bullet}} \right)^2 - \left(\frac{n_{\bullet 2}}{n_{\bullet\bullet}} \right)^2. \end{aligned} \quad (14.12)$$

The reduction of Gini index can be calculated with

$$\Delta Gini = G(t) - P(t_L)G(t_L) - P(t_R)G(t_R). \quad (14.13)$$

If the dependent variable has multiple levels, add corresponding category to the equation. For instance, if there are k levels for independent variable, and the ratios for each category is $p_i (i = 1, 2, \dots, k)$, then

$$G(t_L) = 1 - \sum_i p_i^2. \quad (14.14)$$

c) Pearson chi-square test

The Pearson χ^2 can be calculated with Table 14.5, P value of χ^2 test can be used as the criterion of classification. The smaller the P value, the better the fitness is.

When the response variable is numerical, the decision tree is named as regression tree. F test and variation reduction method can be used as a splitting rule (refer to related materials for more information).

14.5.2.2 *Pruning the tree*

A root node is divided into some daughter nodes, and the daughter nodes are divided into sub-daughter nodes, repeating the process until the tree is full.

The daughter node becomes a terminal node when: (1) the daughter nodes cannot be divided because there is only one subject in each of the daughter nodes; (2) the nodes are pure.

One of the questions that arise in decision tree algorithm is the optimal size of the final tree. A full tree is too large to make reasonable statistical inference due to over fitting the training data and to generalize a new sample. So it is necessary to prune the constructed tree. Pruning is a technique that reduces the size of decision trees by removing sections of the tree that provide little power for classification.

A criterion should be set up before constructing a decision tree, and the splitting will stop when the criterion is met. There are several options, (1) set up an external limit on the number of levels in the maximal tree; (2) Limit a minimum number of individuals in nodes (1% of the total sample size or samples require at least five individuals in nodes); (3) make regulation on the tolerant threshold value of the impurity function.

First, we grow up a full tree, and then prune away the daughter nodes from the last level (i.e. the terminal nodes). There are lots of pruning methods, according to the different methods, a number of child trees were generated as candidates for the appropriately-fit final tree, which was chosen by comparing the qualities of those child trees. No matter the objective is classification or prediction, the quality of a tree depends on the terminal tree while internal nodes only work as the intermediate factors.

14.5.2.3 *Cross validation*

In cross validation, the training sample is randomly split into N (such as 10) subsets with same sample size. $N - 1$ subsets are combined to build a full tree. Then prune this full tree to generate several child-trees. And use the left subset sample to fit each child-tree and calculate the percentage of incorrect classification. Repeat this process for each subset sample and the child-tree with minimum or approximately percentage of incorrect classification will be the final decision tree.

Example 14.3 Department of Obstetrics and Gynecology wants to build a classification tree to discriminate pregnancy outcomes based on the factors including age and alcohol. The outcome variable has two levels: premature delivery or not. 42 pregnant women were recruited and the data were collected, including the outcome (0 = non-premature delivery, 1 = premature delivery), two continuous covariates which are x_1 for age (Yrs) and x_2 for alcohol intake (50 gram/day). The data are shown in Table 14.6.

Solution We select $c = 1.5$ as a cutoff point for independent variable x_2 , and the classification result is shown in Table 14.7.

Table 14.6 Data on premature delivery and age, alcohol intake.

Age (Yrs)	Alcohol intake (50 gram/day)	Premature delivery	Age (Yrs)	Alcohol intake (50 gram/day)	Premature delivery
14	1.2	0	18	1.4	0
16	0.6	0	15	1.7	0
18	0.2	1	15	2.5	0
19	0.7	0	21	1.5	0
20	0.4	0	18	1.9	0
21	1.0	0	23	1.8	1
22	0.8	1	17	2.9	1
24	0.3	0	20	2.6	0
25	0.9	0	23	2.9	0
31	0.8	0	24	2.1	1
29	0.3	0	25	2.5	0
28	0.6	0	28	2.1	1
34	1.0	0	29	1.6	1
36	0.5	0	35	1.7	1
37	1.1	0	32	2.6	1
38	0.7	0	34	2.3	1
39	0.2	0	44	2.1	1
45	0.4	0	37	2.7	1
43	1.0	0	38	2.3	1
45	0.8	0	39	1.6	1
26	1.3	0	42	2.8	1

Notice: The data is adapted from *SPSS and Statistical Analysis*, Chuan Hua Yu as the chief editor. Electronic Industry Press, Beijing, 2007.

Table 14.7 Cross table of nodes and dependent variable.

Daughter nodes	Conditions	Non-premature delivery	Premature delivery	Total
Left daughter node (t_L)	$x_2 < 1.5$	20	2	22
Right daughter node (t_R)	$x_2 \geq 1.5$	7	13	20
Parent node (t)	—	27	15	42

According to Eqs. (14.7)–(14.9), we can get the entropy value and entropy reduction value:

$$E(t_L) = -\frac{20}{22} \ln \left(\frac{20}{22} \right) - \frac{2}{22} \ln \left(\frac{2}{22} \right) = 0.304636,$$

$$E(t_R) = -\frac{7}{20} \ln \left(\frac{7}{20} \right) - \frac{13}{20} \ln \left(\frac{13}{20} \right) = 0.647447,$$

$$E(t) = -\frac{27}{42} \ln \left(\frac{27}{42} \right) - \frac{15}{42} \ln \left(\frac{15}{42} \right) = 0.651757,$$

$$\begin{aligned} \Delta E(t) &= 0.651757 - \left(\frac{22}{42} \right) \times 0.304636 - \left(\frac{19}{42} \right) \times 0.647447 \\ &= 0.183877. \end{aligned}$$

In this example, alcohol intake x_2 is a continuous variable, whose range is 0.2–2.9 (50 gram/day), and there are 25 possible value, so there are 24 possible cutoff points. Similarly, we can calculate the entropy reduction value, Gini reduction index and $-\ln P$ value for each cutoff point. The result shows that the goodness-of-fit is the best with the largest entropy reduction value 0.183877 when alcohol intake x_2 is 1.5 (50 gram/day), while the largest entropy reduction value is 0.031313 when age is equal to 26.5 years old. As the entropy reduction of alcohol intake is larger than that of age, alcohol intake is firstly selected as the independent variable to divide the root node into daughter nodes.

EM (enterprise miner) in SAS can help us to construct a proper decision tree. The procedures include Sample, Explore, Modify, Model, and Assess, which is denoted as SEMMA.

Example 14.4 EM program in SAS was used to analysis the training sample in Example 14.3, and the classification results is shown in Fig. 14.3.

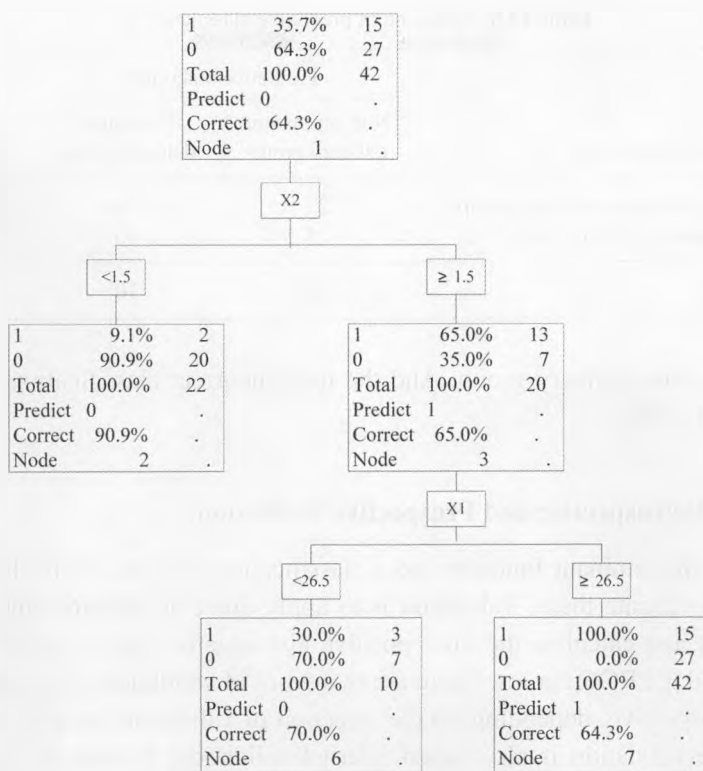


Fig. 14.3 Classification tree.

There are 27 (64.3%) non-premature delivery pregnant women and 15 (35.7%) premature pregnant women in the root node. Based on the alcohol intake, if the alcohol intake $x_2 \geq 1.5$, then it is classified to Node 3; otherwise Node 2. The result shows that Node 2 is a pure node, so that no more splitting is needed; but in Node 3 there are 35.0% of non-premature delivery pregnant women and 65.5% of premature delivery pregnant women, so that more splitting is needed. If age $x_1 \geq 26.5$, the pregnant women would be classify to Node 7; otherwise Node 6. Here Node 7 is pure and Node 6 is up to our pre-set criterion. This tree has three layers.

Table 14.8 shows the cross table of actual observations and predictive outcomes. The last column indicates that 27 non-premature delivery pregnant women were incorrectly classified to non-premature delivery group; ten premature delivery pregnant women were incorrectly classified

Table 14.8 Outcome of predictive classification.

Observed outcomes	Predictive outcomes		Total
	Non-premature delivery group	Premature delivery group	
Non-premature delivery group	27	0	27
Premature delivery group	5	10	15
Total	32	10	42

to premature delivery group. And the total incorrect classification rate is $5/42 = 11.90\%$.

14.6 Retrospective and Prospective Validation

Once a discriminant function and a classification rule are established, we have to validate them. Validation is to apply this classification rule to all samples and calculate the false positive and negative rates, overall accuracy, and the ROC curves. There are two types of validations, retrospective and prospective, depending on the selection of validation samples. Retrospective validation is also called internal validation. It used the training samples to estimate the false positive rate, false negative rate and overall misclassification rate. However, such validation tends to underestimate the true errors. For practical use, prospective validation is necessary. We will use the rule in clinical applications only when we are satisfied with the performance of the classification for the prospective samples. Therefore, to establish a classification rule, we need both training samples and validation samples. Validation samples should also be classified according to the gold standard. They should be collected with care, including the quality assurance for the accuracy of data and large enough sample sizes. In clinical research, we can divide collected data into two groups, one for training and one for validation. Samples can be randomly selected into one of the two groups.

Besides calculating the error rates using prospective validation, we can use jackknife or cross-validation techniques to estimate the error rates. Suppose that we have n samples in our training data, we can take the first sample out and use the rest $n - 1$ samples to develop the classification

rule. We then apply the classification rule to the first sample and determine whether we have the correct classifications. We then rotate the sample to be validated by the classification rule of the rest samples. These classification rules may be different because of the exclusion of different samples. After we exhaust all possible exclusion of samples, we can calculate the prediction errors based on the prediction of every sample. This method is the jackknife validation or leaving one out (LOO) validation. It approximates the prospective uses of the classification rule to new data.

Jackknife can still be conservative. If sample size is reasonable, we can group data into k subgroups. We then evaluate prediction results of one of the k groups based on the classification rule of the rest $k - 1$ groups. After altered all k groups as the predicted dataset, we can evaluate overall prediction errors. Usually k is in the range of 5 to 20, depending on the sample sizes. This method is called k -fold cross-validation method. It is usually a better approximation to prospective validation than jackknife validation.

Example 14.5 We examined the performance of the classification rule in Example 14.1 using three different approaches: retrospective validation, jackknife validation, and prospective validation using 31 newly collected samples. The results are given in Tables 14.9–14.11. Among them, the retrospective validation gave the most optimistic picture (lowest classification error), which demonstrated its potential bias of underestimating classification errors. The total error rates are similar, but the rates for individual groups are quite different between the jackknife and the prospective classifications. When sample size of training data is large enough, the difference between jackknife and prospective validations is negligible.

Table 14.9 Classification error from retrospective validation.

Original group	Classification by discriminant analysis			Total	Error rate (%)
	A_1	A_2	A_3		
A_1	62	4	2	68	8.82
A_2	1	41	1	43	4.65
A_3	1	0	19	20	5.02
Total	64	45	22	131	6.87

Table 14.10 Classification error from jackknife validation.

Original group	Classification by discriminant analysis			Total	Error rate (%)
	A_1	A_2	A_3		
A_1	60	6	2	68	11.76
A_2	2	40	1	43	6.98
A_3	1	0	19	20	5.00
Total	63	46	22	131	9.16

Table 14.11 Classification error from prospective validation.

Original group	Classification by discriminant analysis			Total	Error rate (%)
	A_1	A_2	A_3		
A_1	14	1	0	15	6.67
A_2	1	9	1	11	18.18
A_3	0	0	5	5	0.00
Total	15	10	6	31	9.68

14.7 Considerations in Applications

14.7.1 *The quality of training data is the key for a good discriminant analysis*

The classification rule developed by a discriminant analysis can be trusted if the classifications of the training data are correct, the explanatory variables are appropriately selected, the values of explanatory variables are accurately measured, and the sample size of training data is large enough. Using inaccurate and unreliable information will inevitably result in unreliable classification rules: Garbage in, garbage out!

14.7.2 *The more groups, the less accurate of the classification rules*

When some variables good in separating two groups are included for discriminating more than two groups, the efficiency might decrease. At this time, we may want to perform multiple binary classifications conditioning

on the previously used classification results. The recursive partitioning algorithm, also known as classification and regression tree (CART) technique, uses the conditional consecutive binary classifications to discriminate multiple groups.

14.7.3 *Fisher's and Bayesian discriminations are the same for two-class classification*

For two-class discriminant analysis, we can derive a discriminant function based on the difference between two Bayesian classification functions. When the difference is used, the Bayesian classification function is equivalent to the Fisher's discriminant function.

14.7.4 *Decision tree*

The foothold of this method is to divide the parent node such as root node into two daughter nodes. Appropriate partition rules and optimal cutoff points can have certain influence on the effect of decision trees. Compared to the traditional methods, decision trees have definite merits in various aspects. On one hand, it can tackle all kinds of data and no particular requirements are made for the distribution of the data. On the other hand, the hierarchical structure of the decision tree is clear as variables are analyzed in the order of their relative importance. And simultaneously, the fitting effect of the tree can be evaluated. However, selecting the optimal decision index and the process of pruning the tree are so complicated that the application is relatively restricted.

14.7.5 *Prospective validation*

All classification rules developed through discriminant analysis should be validated prospectively before their use in clinical practice. Some researches failed to use the prospective validation because they cannot collect enough samples. Even though the retrospective validation shows good performance, the outcomes cannot guarantee their future performance. This is like testing students with problems that answers were given. Even though the students have the perfect results, it does not indicate that students know the materials well.

14.8 Computerized Experiments

Data of 131 diabetic patients with measurements of 11 explanatory variables and the 31 diabetic patients for prospective validation are stored in the file EYE1.DAT and EYE2.DAT respectively in the Example 11 of Appendix IV. Using these two data files, you are asked to perform the following tasks.

Experiment 14.1 Conducting a stepwise discriminant analysis with both P1 and P2 (entrance and staying significant levels) of 0.05.

Commands from line 01 to line 03 in Program 14.1 illustrates how to input data for training samples. Line 07 calls the SAS procedure PROC STEPDISC to analyze the EYE1 data. The method used to select variables is STEPWISE and significance levels of being selected into the model and of staying in the model are all 0.05. Line 08 lists all explanatory variables available to build the discriminant function. Line 09 specifies the variable of the classification using the gold standard. Line 10 tells computer to run the program. The results of this program should select five variables: age, vision, at, bv, and qpv.

Program 14.1 Stepwise discriminant analysis.

Line	Program
01	DATA EYE1;
02	INPUT AGE TIME GLUCOSE VISION AT AV BT BV QPT QPV GROUP \$;
03	CARDS;
04	49 2 191 1.5 12.25 235.4 52.5 417.57 78.5 27.43 A1

05	60 1 134 0.4 15.5 393.64 54 541.13 76.5 17.96 A3
06	;
07	PROC STEPDISC DATA=EYE1 METHOD=STEPWISE SLENTRY=0.05 SLSTAY=0.05;
08	VAR AGE TIME GLUCOSE VISION AT AV BT BV QPT QPV;
09	CLASS GROUP;
10	RUN;

Experiment 14.2 Conducting a validation of the classification rule of Experiment 14.1. Commands from line 01 to line 03 of Program 14.2 are

used to input data and set up data file EYE2.DAT for validation samples. Data for training samples have been set up in Program 14.1. Line 04 calls the SAS procedure DISCRIM to perform discriminant analysis. In this procedure, we select DATA=EYE1, the training data, and TESTDATA=EYE2, the validation data. Option LIST requests the display of the retrospective validation results. Option CROSSLIST requests the display of jackknife results. Finally, option TESTLIST requests the display of prospective validation results. If we are interested only in summary of the numbers of correct and incorrect classifications, but not individual patient results, we can change the above options to LISTERR, CROSSLISTERR, and TESTLISTERR. If we are interested only in the false positive and false negative rates as well as the misclassification rate, we need only to call CROSSVALIDATE to calculate these results of jackknife validation. Line 05 defines the class variable. Line 06 gives the list of explanatory variables to be used in the model. Line 07 submits the program.

Program 14.2 Validation of the classification rule.

Line	Program
01	DATA EYE2;
02	INPUT AGE TIME GLUCOSE VISION AT AV BT BV QPT QPV GROUP \$;
03	CARDS;
04	54 10 137 0.7 13.75 275.94 55.5 492.3 77.5 35.32 A1

05	52 18 296 0.5 15.25 258.36 51.25 439.06 80.5 7.89 A3
06	;
07	PROC DISCRIM DATA=EYE1 TESTDATA=EYE2 LIST CROSSLIST TESTLIST;
08	CLASS GROUP;
09	VAR AGE VISION AT BV QPV;
10	RUN;

Experiment 14.3 Classification tree building The data in Example 14.3 is in Appendix V. Save the data in file A.DAT as training sample. First run program 14.3 to establish dataset A, and then use Enterprise Miner in SAS to build the classification tree (see Program 14.3).

Program 14.3 Classification tree building (establish dataset).

Line	Program					
01	DATA A;					
02	INPUT x1 x2 y@@;					
03	CARDS;					
04	14	1.2	0	18	1.4	0
05	16	0.6	0	15	1.7	0
06					
07	26	1.3	0	42	2.8	1
	RUN;					

Solution

analysis
enterprise miner

Open EM in SAS, establish the decision tree by running the module, the process is shown as follows:



14.9 Practice and Experiments

1. Use the data in Example 14.1 (EYE1.DAT in Lab Experiment) to classify mild and severe patients.
- (1) Using stepwise discriminant analysis to select significant explanatory variables (both P1 and P2 are 0.05).
- (2) Using the selected variables to develop two classification functions (using the ratio of sample size of two groups as the prior probability).

- (3) Assigning a patient into the group that has larger value of the classification functions. Using this rule to perform retrospective validation (or internal validation).
- (4) Taking the differences between two classification functions to generate a new discriminant function. How to use the discriminant function to classify patients such that the results are the same as (c).

2. (Cont'd from 1) Define a dependent variable Y . When a patient has mild disease, set Y = the number of mild patients/the number of mild and severe patients, or the proportion of mild patients among mild and severe patients. When a patient has severe disease, set Y = the proportion of patients with severe disease among mild and severe patients.

- (1) Using the stepwise regression procedure in the previous chapter to select significant variables in the regression equation (using proper thresholds for F_{in} and F_{out}). Compare the results with the above discriminant analysis.
- (2) Compare the regression coefficients with the coefficients in the discriminant function. Are they similar or different?

3. Perform similar discriminant analysis as in Exercise 1 and stepwise regression analysis as in Exercise 2 for two groups of mild and moderate, and two groups of moderate and severe. Compare the results in Example 14.1 with all three two-group comparisons. Are they similar or different and how?

(1st edn. Ying Lu; 2nd edn. Jinxin Zhang, Jibin Li, Jiqian Fang)



Chapter 15

Logistic Regression

Multiple regression is used to analyze the relationship of a dependent variable with several independent variables. The purposes of analysis include adjustment of confounding factors, selection of significant covariates related to the dependent variable and prediction of the value of dependent variable. In multiple linear regression, Y is a continuous random variable and it is usually required to have a normal distribution given the values of independent variables. In medical practice, people often need to deal with the situation in which the dependent variable only has two possible values, such as ill and healthy, death and alive, recover and not recover. In this situation, logistic regression is considered.

15.1 Logistic Regression Model

15.1.1 *Basic concept of logistic regression model*

Example 15.1 (*cross-sectional study*) In a study of risk factors related to emergency treatment of acute myocardial infarction (AMI), in five years 200 AMI cases from a hospital were collected with their disease history and treatments recorded. Now define the dependent variable Y , $Y = 0$ means success (i.e. survive) and $Y = 1$ means failure (i.e. death). Three risk factors were analyzed. They are X_1 ($X_1 = 1$ refers to shock before rescue and $X_1 = 0$ otherwise), X_2 ($X_2 = 1$ refers to heart failure before rescue and $X_2 = 0$ otherwise) and X_3 ($X_3 = 1$ refers to no treatment in 12 hours after AMI and $X_3 = 0$ otherwise). The data of risk factors and treatment results were collected (see Table 15.1). By regarding the 200 AMI cases as a random sample of all AMI cases, this study was subject to a cross-sectional study.

Table 15.1 The outcomes and related risk factors of 200 AMI cases.

$Y = 0$				$Y = 1$			
X_1	X_2	X_3	N	X_1	X_2	X_3	N
0	0	0	35	0	0	0	4
0	0	1	34	0	0	1	10
0	1	0	17	0	1	0	4
0	1	1	19	0	1	1	15
1	0	0	17	1	0	0	6
1	0	1	6	1	0	1	9
1	1	0	6	1	1	0	6
1	1	1	6	1	1	1	6

The purpose of this study is to analyze the risk factors related to the failure of emergency treatment and to build up a model to predict the probability of failure, P . If a linear regression is used, the model will be

$$P = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

and the regression equation estimated by sample will be

$$\hat{P} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3.$$

In the linear regression model, X_1, X_2, X_3 may take any values in the interval $(-\infty, \infty)$ and the estimates of $\beta_0, \beta_1, \beta_2, \beta_3$ are not constrained. Thus, there is no guarantee that the value of \hat{P} should fall in the interval $[0, 1]$. But a probability cannot be bigger than 1 or less than 0, if it is, the result cannot be explained in practice. In order to have a restricted range of $[0, 1]$ for the probability \hat{P} , it is suggested to employ a transformation (see Fig. 15.1)

$$\ln \left(\frac{P}{1 - P} \right) = W. \quad (15.1)$$

Here $P/(1 - P)$ is so-called odds, the left side of (15.1) is the logarithm of odds, called "logit of P ". Then set up a linear regression model for W , that is,

$$W = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

or

$$\ln \left(\frac{P}{1-P} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad (15.2a)$$

or

$$\text{logit}(P) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3. \quad (15.2b)$$

Here the log(odds) or logit of P is a linear function of the explanatory variables X_1, X_2, X_3 . To express the relationship between P and X_1, X_2, X_3 directly, (15.2a) can be rewritten as

$$P = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}} \quad (15.2c)$$

or

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)}}. \quad (15.2d)$$

In general, if there are k explanatory variables X_1, X_2, \dots, X_k , the model of logistic regression can be expressed with the following four equivalent forms:

$$\ln \left(\frac{P}{1-P} \right) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k, \quad (15.3a)$$

$$\text{logit}(P) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k, \quad (15.3b)$$

$$P = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k}}, \quad (15.3c)$$

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}}. \quad (15.3d)$$

15.1.2 Parameter estimation of logistic regression

The maximum likelihood method is often used for parameter estimation of logistic regression.

Likelihood function is defined as the probability of that a random trial results in the current situation in theory, which depends on the parameters of the assumed theoretical model. For Example 15.1 according to the model

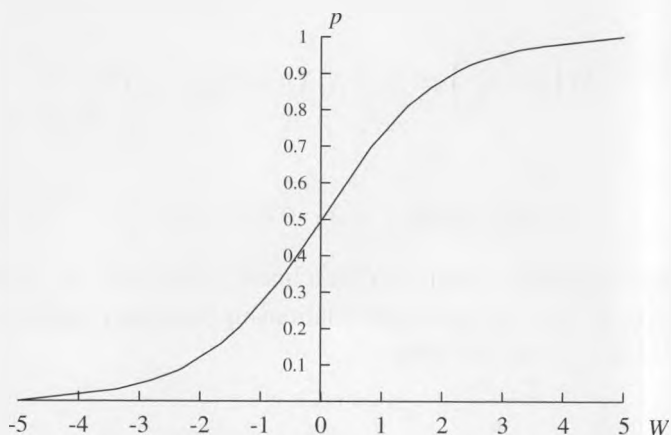


Fig. 15.1 The relationship between P and W in Eq. (15.1).

(15.2c), given the values of (x_{i1}, x_{i2}, x_{i3}) , the i th individual's probabilities of failure and success are

$$P(Y_i = 1) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}}$$

and

$$P(Y_i = 0) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}}$$

respectively. The two equations can be pooled as

$$P(Y_i) = \left[\frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}} \right]^{Y_i} \left[\frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}} \right]^{1-Y_i}$$

Then the likelihood function and the log-likelihood function of Example 15.1 can be written as

$$Lik(\beta_0, \beta_1, \beta_2, \beta_3)$$

$$= \prod_{i=1}^{200} \left[\frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}} \right]^{Y_i} \left[\frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}} \right]^{1-Y_i}$$

Solution The values of explanatory variables of patient A are $X_1 = 0$, $X_2 = 1$ and $X_3 = 0$. By Eq. (15.4d)

$$P_A = \frac{1}{1 + e^{-(-2.0858 + 1.1098X_1 + 0.7028X_2 + 0.9751X_3)}}$$

$$= \frac{1}{1 + e^{-(-2.0858 + 0.7028)}} = 0.200.$$

Those of patient B are $X_1 = 1$, $X_2 = 1$ and $X_3 = 1$. By Eq. (15.4d)

$$P_B = \frac{1}{1 + e^{-(-2.0858 + 1.1098X_1 + 0.7028X_2 + 0.9751X_3)}}$$

$$= \frac{1}{1 + e^{-(-2.0858 + 1.1098 + 0.7028 + 0.9751)}} = 0.669.$$

15.1.3 The hypothesis testing for logistic model

After the estimation of regression coefficients, it is necessary to test the significance of model and if the population coefficient is zero.

15.1.3.1 The test for goodness-of-fit

Usually the value of log-likelihood function shows the goodness-of-fit of the model. The larger the value of log-likelihood function, the better the fit.

The hypotheses to be tested are

H_0 : The model fits the data

H_1 : The model does not fit the data

Denote the maximum log-likelihood with $\ln L$. It can be proved, for a large sample, when H_0 is true, $-2 \ln L$ follows a chi-square distribution with degrees of freedom $\nu = N - k - 1$. Here N is the sample size, k is the number of explanatory variables in model.

When $-2 \ln L \approx \chi^2 > \chi^2_{\alpha, \nu}$, the null hypothesis H_0 is rejected. It means the model does not fit the observed data and is not suitable to use for prediction. For example, the third column in Table 15.3 lists the values of $-2 \ln L$ for four models. The $-2 \ln L$ of model 1 is 244.246, $\nu = 200 - 0 - 1 = 199$, $P = 0.02$, H_0 is rejected, which means model 1 does not fit the data. While for model 4, $P = 0.09$, H_0 cannot be rejected, which means model 4 fits the data.

15.1.3.2 The likelihood ratio test

This test could be used to test whether the fitness of two models are the same. The hypotheses to be tested are

H_0 : The two models are the same in goodness-of-fit

H_1 : The two models are different in goodness-of-fit

The statistic being used to describe the difference in terms of fitness between the two models is

$$G = -2 \ln L - (-2 \ln L') = -2 \ln \left(\frac{L}{L'} \right). \quad (15.5)$$

Here $-2 \ln L$ refers to model 1 with k_1 variables, $-2 \ln L'$ refers to model 2 with k_2 variables, $k_1 < k_2$.

It can be proved, for large sample, when H_0 is true, G follows a chi-square distribution with degrees of freedom $\nu = k_2 - k_1$.

Given the value of α , if the P -value is less than α , then H_0 is rejected, which means the two models are significantly different in terms of fitness so that it is worthwhile to take the model with more variables. Otherwise, H_0 is not rejected, which means the model with less variables is acceptable.

In Table 15.3, comparing to model 1, the model 2 has one more variable X_1 , $-2 \ln L = 244.346$, $-2 \ln L' = 236.736$, $G = 7.610$, $\nu = 1 - 0 = 1$, $P < 0.01$. It means the goodness-of-fit is significantly improved when including X_1 into the model and model 2 has a better prediction effect than model 1. By the same procedure, one can see that including X_2 and X_3 also improves the goodness-of-fit significantly so that model 4 is the best model for predicting the treatment result of AMI.

15.1.3.3 The score test

The hypotheses to be tested are

H_0 : Some of the coefficients equal zero

H_1 : These coefficients do not equal zero

Table 15.2 Results of parameter estimation and Wald test for Example 15.1.

Variable	$\hat{\beta}$	SE($\hat{\beta}$)	Wald χ^2	P-value	OR
Intercept	-2.0858	0.3513	35.2632	0.0001	
X_1	1.1098	0.3485	10.1422	0.0014	3.034
X_2	0.7028	0.3292	4.5587	0.0328	2.019
X_3	0.9751	0.3440	8.0365	0.0046	2.651

Table 15.3 The models fitted for the data in Example 15.1.

Model	Parameter	$-2\ln L$	G	Score
1	β_0	244.346	—	—
2	β_0, β_1	236.736	7.610	7.854
3	$\beta_0, \beta_1, \beta_2$	227.200	9.536	6.898
4	$\beta_0, \beta_1, \beta_2, \beta_3$	222.616	4.583	5.309

After setting up a likelihood function or logarithmic likelihood function, one could use the statistical software to find the values of parameters which maximize the likelihood function or log-likelihood function. Based on the data in Example 15.1, the results of logistic regression are listed in Tables 15.2 and 15.3.

In Table 15.2, column 2 gives the estimates of the intercept and coefficients in the model of logistic regression,

$$\hat{\beta}_0 = -2.0858, \hat{\beta}_1 = 1.1098, \hat{\beta}_2 = 0.7028, \hat{\beta}_3 = 0.9751.$$

Therefore, the logistic regression equation can be expressed as any of the following four equations:

$$\ln \left(\frac{P}{1-P} \right) = -2.0858 + 1.1098X_1 + 0.7028X_2 + 0.9751X_3, \quad (15.4a)$$

$$\text{Logit}(P) = -2.0858 + 1.1098X_1 + 0.7028X_2 + 0.9751X_3, \quad (15.4b)$$

$$P = \frac{e^{-2.0858 + 1.1098X_1 + 0.7028X_2 + 0.9751X_3}}{1 + e^{-2.0858 + 1.1098X_1 + 0.7028X_2 + 0.9751X_3}}, \quad (15.4c)$$

$$P = \frac{1}{1 + e^{-(-2.0858 + 1.1098X_1 + 0.7028X_2 + 0.9751X_3)}}. \quad (15.4d)$$

From (15.4a), one can see that $\hat{\beta}_0 = -2.0858$ is the $\ln(\text{odds})$ when all the independent variables take values zero;

$$e^{\hat{\beta}_0} = e^{-2.0858} = 0.1242$$

is the odds of failure for the individual without shock, heart failure and delay before treatment. When the probability of failure is very low and the probability of success closes to 1, this value closes to failure rate.

$$\begin{aligned} \hat{\beta}_1 &\approx \ln(\text{odds}_{\text{with shock}}) - \ln(\text{odds}_{\text{without shock}}) \\ &= \ln\left(\frac{\text{odds}_{\text{with shock}}}{\text{odds}_{\text{without shock}}}\right) = 1.1098. \end{aligned}$$

Therefore,

$$e^{\hat{\beta}_1} = e^{1.1098} = 3.0338$$

is the “odds ratio” between with shock and without shock before treatment when other conditions (heart failure and delay of treatment) are kept the same. Odds ratio is an important measurement introduced in Chap. 2. When the failure rate is low, it closes to a relative risk of failure for patients with shock before treatment comparing to those without shock.

In general, when P refers to the probability of failure, if the regression coefficient is positive $\hat{\beta} > 0$, hence $e^{\hat{\beta}} > 1$, then the corresponding variable is a risk factor; otherwise, it is a protective factor. In this example, the coefficients of three variables are all positive so that shock, heart failure and delay before treatment are all risk factors.

A logistic model may be used for forecasting.

Example 15.2 Patient *A* with heart failure and without shock was sent to the hospital in five hours after AMI symptom appearing. Patient *B* with heart failure and shock was sent to hospital 18 hours after AMI symptom. Estimate their probabilities of failure respectively.

As the first step, we work out a logistic regression under the condition of H_0 , and then calculate a statistic, which is called score,

$$\text{Score} = \mathbf{S}'(\mathbf{COV})\mathbf{S}. \quad (15.6)$$

Here \mathbf{S} is a vector consists of all the first order partial derivative of log-likelihood function on the estimators of the parameters, \mathbf{COV} is the variance-covariance matrix of the estimators of the parameters. When sample size is large enough, this statistic (score) approximately follows a chi-square distribution with a degree of freedom equal to the difference between the numbers of estimated coefficients.

For example, to test

$$H_0 : \beta_0 \neq 0, \beta_1 \neq 0 \text{ but } \beta_2 = 0 \quad (\text{model 2})$$

$$H_1 : \beta_0 \neq 0, \beta_1 \neq 0 \text{ and } \beta_2 \neq 0 \quad (\text{model 3})$$

the data are fitted under H_0 (model 2), and then calculate the statistic, $\text{Score} = 6.898$, $\nu = 1$, $P < 0.01$ so that H_0 is rejected, which means model 3 is worthwhile to try. Again to test

$$H_0 : \beta_0 \neq 0, \beta_1 \neq 0, \beta_2 \neq 0 \text{ but } \beta_3 = 0 \quad (\text{model 3})$$

$$H_1 : \beta_0 \neq 0, \beta_1 \neq 0, \beta_2 \neq 0 \text{ and } \beta_3 \neq 0 \quad (\text{model 4})$$

the data are fitted under H_0 (model 3), and then calculate the statistic, $\text{Score} = 5.309$, $\nu = 1$, $P < 0.05$ so that H_0 is rejected, which means model 4 is worthwhile to try.

One can see that the advantage of the score test is to fit a model with less parameters and decide whether the complicated model is needed. Obviously, it is useful in forward selection of variables.

15.1.3.4 Wald test

The hypotheses to be tested are also

$$H_0 : \text{Some of the coefficients equal zero}$$

$$H_1 : \text{These coefficients do not equal zero}$$

In particular, the most popular situation is to test for whether one of the coefficients equals zero or not, that is

$$H_0 : \beta_i = 0 \quad H_1 : \beta_i \neq 0$$

Different from the score test, as the first step, we work out a logistic regression under the condition of H_1 , and then calculate a statistic, which is called Wald χ^2 ,

$$\chi_i^2 = \left[\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \right]^2. \quad (15.7)$$

Here, $\hat{\beta}_i$ is an estimated value of the regression coefficient β_i . SE is the standard error of $\hat{\beta}_i$. When H_0 is true, if the sample size is large enough, then the statistic approximately follows a chi-square distribution with one degree of freedom. In Table 15.2, the Wald χ^2 for X_1 (shock) is

$$\chi^2 = \left[\frac{1.1098}{0.3485} \right]^2 = 10.1422.$$

$P < 0.01$, H_0 is rejected so that the population coefficient for X_1 is not equal to zero.

One can see that the advantage of the Wald test is to easily find the insignificant variables after fitting a model with all the variables. Obviously, it is useful in backward selection of variables.

It can be proved, under H_0 the three test statistics of likelihood ratio test, score test and Wald test are asymptotically equivalent. For large sample size n , the values of the three test statistics will tend to be close to each other. For medium or small sample size, their differences can be more serious, the results may not be consistent. In practice, among the three, likelihood ratio test is preferred and its result is more robust. When sample size is not large, the distribution of score is closer to chi-square distribution so that it is more sensitive. Wald test is easy to calculate but more conservative.

15.1.4 More examples

Example 15.3 (cohort study) In order to study the relationship between coronary heart disease (CHD) and endogenous catecholamine (CAT), two populations with high CAT and low CAT were followed up respectively

Table 15.4 Data on coronary heart disease and endogenous catecholamine.

Stratum	CAT = 1 (high)		CAT = 1 (low)	
	Case	Normal	Case	Normal
AGE < 55, ECG = 0	1	17	7	257
AGE < 55, ECG = 1	3	7	14	52
AGE ≥ 55, ECG = 0	9	15	30	107
AGE ≥ 55, ECG = 1	14	5	44	27
Total	27	44	95	443

Table 15.5 Results of logistic regression for the data in Table 15.4.

Model*	Variable	$\hat{\beta}$	$SE(\hat{\beta})$	Wald χ^2	P-value	OR
1	Constant	-1.5396	0.1131	185.4328	0.0000	—
	CAT	1.0512	0.2693	15.2331	0.0001	2.8612
2	Constant	-3.3725	0.2588	169.8494	0.0000	—
	CAT	0.6390	0.2621	4.2735	0.0387	1.9328
	AGE	2.0550	0.2430	61.4541	0.0000	7.8072
	ECG	1.8785	0.2588	59.7407	0.0000	6.5440

*For model 1, $-2\ln L = 595.913$; for model 2, $-2\ln L = 445.920$

during a seven-year period. The incident cases of CHD occurred in two populations were counted. Considering the confounding of age and abnormal of electrocardiogram (ECG), the data were stratified by these two factors in Table 15.4.

This is a cohort study. The outcome is described by a binary variable Y , $Y = 1$ refers to suffering from CHD and $Y = 0$ otherwise. A logistic regression model for $P(Y = 1)$ on CAT, age and ECG has been constructed. The results are listed in Table 15.5.

The results on model 1 show that the odds ratio related to CAT is $e^{1.0512} = 2.8612$. It means that people with high CAT have almost 3 times higher risk of CHD comparing to people with low CAT before adjusting for age and ECG. When age and ECG are adjusted in model 2, the OR becomes $e^{0.639} = 1.9328$, suggesting that the higher OR before adjusting is partly due to the confounding effect of age and ECG. Comparing to model 1, model 2

Table 15.6 Data on salted vegetable intake and esophageal cancer.

Age group	Case		Control	
	Intake	No intake	Intake	No intake
1 (25–34)	1	0	8	98
2 (35–44)	4	6	24	186
3 (45–54)	25	22	32	148
4 (55–64)	56	38	28	139
5 (65–74)	19	36	18	88
6 (75+)	5	8	0	31
Total	110	110	110	690

has a significant improvement of goodness-of-fit, $G = 149.993$ and $\nu = 2$, $P < 0.001$. The final model is

$$\text{Logit}(P) = -3.3725 + 0.6390\text{CAT} + 2.0550\text{AGE} + 1.8785\text{ECG}.$$

This model may be used for predicting the risk of CHD.

Example 15.4 (group case-control study) The data of a case-control study on esophageal cancer and salted vegetable intake are given in Table 15.6. Analyze the relationship between esophageal cancer and salted vegetable intake by adjusting confounding effect of age.

In the case-control study, research purpose is to analyze the relationship between disease and risk factors. Odds ratio is a common measure of this relationship. If the confounding effect of age is not considered, the *OR* of esophageal cancer related with salted vegetable intake may be calculated directly,

$$OR = \frac{110 \times 690}{110 \times 110} = 6.273. \tag{15.8}$$

If the confounding effect of age needs to consider, the *OR* should be estimated by stratification method. However, in Table 15.6, the numbers in some strata are very small and even there is zero in the strata age 25–34 and 75+ so that the *OR* cannot be well estimated directly stratum by stratum. The logistic regression model may solve such kind of problems and utilize the information from all strata sufficiently. It is convenient to adjust for several confounding factors and to estimate the adjusted *OR*.

Let us use a binary variable X for salted vegetable intake that $X = 1$ refers to salted vegetable intake and $X = 0$ otherwise. The results of logistic regression on salted vegetable intake only are

$$\text{logit}(P) = -1.8362 + 1.8362X, \quad (15.9)$$

$$OR = e^{1.8362} = 6.273. \quad (15.10)$$

Note that the result in (15.10) is exactly the same as that in (15.8) because both are the unadjusted OR .

If age is considered as a categorical variable, five binary variables $Age2$, $Age3$, $Age4$, $Age5$, $Age6$ may be used to present age group 35–44, 45–54, 55–64, 65–74 and 75+. Say, if one's age is in group 35–44, his $Age2 = 1$ and all others = 0; if one's age is in group 25–34, all five dummy variables equal zero. Taking these dummy variables into the logistic model, the results are

$$\begin{aligned} \text{logit}(P) = & -4.9806 + 1.5186 Age2 + 3.0502 Age3 + 3.7888 Age4 \\ & + 3.8808 Age5 + 3.8822 Age6 + 1.7163X, \end{aligned} \quad (15.11)$$

$$OR = e^{1.7163} = 5.564. \quad (15.12)$$

It shows that after adjusting for age, the OR is lower than that in (15.8) and (15.10), suggesting the existence of a little confounding effect of age.

It needs to pay attention on that the ratio between the sample sizes for cases and controls does not represent the actually ratio between the numbers of patients and healthy persons in general population so that the intercept given by the regression is not the estimated prevalent rate in the population.

In fact, from Eq. (15.9), $Intercept = -1.8362$. We have

$$e^{-1.8362} = 0.1594 \approx \frac{110}{690}. \quad (15.13)$$

The left-hand side is the "odds ratio" when $X = 0$, while the right-hand side is the ratio between the number of cases and number of controls in the data set of Table 15.6.

Again, from Eq. (15.11), $Intercept = -4.9806$. We have

$$e^{-4.9806} = 0.0069 \approx \frac{0}{98}. \quad (15.14)$$

Now the left-hand side is the "odds ratio" when $X = 0$ and $\text{Age2} = \text{Age3} = \text{Age4} = \text{Age5} = \text{Age6} = 0$, while the right-hand side is the ratio between the number of cases and number of controls in the age group 25–34 of Table 15.6.

Obviously, neither (15.13) nor (15.14) reflects the real situation in the population. Therefore in the logistic model for case-control study, the intercept in the regression equation reflects the situation in the case-control sample only, it does not make sense to the population and this equation cannot be used for prediction. The explanation of the resulted logistic equation would be focused on the regression coefficients.

If one really wishes to use a logistic model from a case-control study for prediction, the real prevalence rate P_0 in the population should be known and the intercept $\hat{\beta}_0$ in the regression equation should be replaced with the adjusted intercept $\hat{\beta}'_0$

$$\hat{\beta}'_0 = \ln \left(\frac{P_0}{1 - P_0} \right) - (\hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 + \cdots + \hat{\beta}_k \bar{X}_k). \quad (15.15)$$

Here $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$ are the sample means of X_1, X_2, \dots, X_k .

For Eq. (15.9), suppose in a population of age over 30 the prevalent rate of esophageal cancer is 0.00058, the alcohol drinking rate is 0.015 ($= \bar{X}$), the adjusted intercept is

$$\begin{aligned} \hat{\beta}'_0 &= \ln \left(\frac{P_0}{1 - P_0} \right) - (\hat{\beta} \bar{X}) \\ &= \ln \left(\frac{0.00058}{1 - 0.00058} \right) - (1.8362 \times 0.015) = -7.4547. \end{aligned}$$

Replacing the intercept in Eq. (15.9) with this value, the model could be used for prediction is any of the following four equations:

$$\begin{aligned} \log \text{it}(P) &= -7.4547 + 1.8362X, \\ \ln \left(\frac{P}{1 - P} \right) &= -7.4547 + 1.8362X, \\ P &= \frac{e^{-7.4547 + 1.8362X}}{1 + e^{-7.4547 + 1.8362X}}, \quad P = \frac{1}{1 + e^{-(-7.4547 + 1.8362X)}}. \end{aligned}$$

Table 15.7 Values of three daily life factors.

Variables	Values
Diet habit (X_1)	0,1,2,3,4 (0 = good, 4 = poor)
Salted food (X_2)	0,1,2,3,4 (0 = non, 4 = frequently)
Mental status (X_3)	0 = poor, 1 = good

Table 15.8 Data of 50 pairs of stomach cancer cases and controls.

No.	Case			Control		
	X_1	X_2	X_3	X_1	X_2	X_3
1	2	4	0	3	1	0
2	3	2	1	0	1	0
3	3	0	0	2	0	1
4	3	0	0	2	0	1
5	3	0	1	0	0	0
...
49	1	2	1	0	0	1
50	2	0	1	0	3	1

(The complete data will be given in Experiment 15.2)

15.2 Conditional Logistic Regression

Example 15.5 (1:1 matched case-control study) The relationship of stomach cancer and three daily life factors was studied in a city. The study chose a paired design. A healthy control was chosen according to the gender, age and living place for each case. The daily life factors and their values for different categories were listed in Table 15.7. In total 50 pairs of case and control were recorded. The data are listed in Table 15.8.

In this study, the case is comparable with his control in each pair in the sense that they have same gender, age and living place. However, the case is not comparable with the controls in other pairs. Therefore, the logistic model is built on the basis of the disease status and exposure in each pair of subjects.

In any pair, the case is indicated by A and the control is indicated by B , $Y = 1$ refers to suffering disease and $Y = 0$ otherwise. The probability of

only one within the pair suffering from the disease is

$$\begin{aligned} P(\text{either } Y_A = 1 \text{ or } Y_B = 1) \\ = P(Y_A = 1)P(Y_B = 0) + P(Y_A = 0)P(Y_B = 1). \end{aligned} \quad (15.16)$$

Under the condition that only one within the pair suffering from the disease, the probability of A suffering from the disease is

$$\begin{aligned} P(Y_A = 1 | \text{either } Y_A = 1 \text{ or } Y_B = 1) \\ = \frac{P(Y_A = 1)P(Y_B = 0)}{P(Y_A = 1)P(Y_B = 0) + P(Y_A = 0)P(Y_B = 1)}. \end{aligned} \quad (15.17)$$

Let

$$P(Y_A = 1) = \frac{e^{(\beta_0 + \beta_1 X_1^A + \beta_2 X_2^A + \dots + \beta_k X_k^A)}}{1 + e^{(\beta_0 + \beta_1 X_1^A + \beta_2 X_2^A + \dots + \beta_k X_k^A)}}, \quad (15.18a)$$

$$P(Y_B = 1) = \frac{e^{(\beta_0 + \beta_1 X_1^B + \beta_2 X_2^B + \dots + \beta_k X_k^B)}}{1 + e^{(\beta_0 + \beta_1 X_1^B + \beta_2 X_2^B + \dots + \beta_k X_k^B)}}, \quad (15.18b)$$

$$P(Y_A = 0) = \frac{1}{1 + e^{(\beta_0 + \beta_1 X_1^A + \beta_2 X_2^A + \dots + \beta_k X_k^A)}}, \quad (15.18c)$$

$$P(Y_B = 0) = \frac{1}{1 + e^{(\beta_0 + \beta_1 X_1^B + \beta_2 X_2^B + \dots + \beta_k X_k^B)}}. \quad (15.18d)$$

Substitute the above four equations into Eq. (15.17), and simplify it as

$$\begin{aligned} P(Y_A = 1 | \text{either } Y_A = 1 \text{ or } Y_B = 1) \\ = \frac{e^{(\beta_1 X_1^A + \beta_2 X_2^A + \dots + \beta_k X_k^A)}}{e^{(\beta_1 X_1^A + \beta_2 X_2^A + \dots + \beta_k X_k^A)} + e^{(\beta_1 X_1^B + \beta_2 X_2^B + \dots + \beta_k X_k^B)}} \end{aligned} \quad (15.19a)$$

or

$$\begin{aligned} P(Y_A = 1 | \text{either } Y_A = 1 \text{ or } Y_B = 1) \\ = \frac{1}{1 + e^{[\beta_1 (X_1^A - X_1^B) + \beta_2 (X_2^A - X_2^B) + \dots + \beta_k (X_k^A - X_k^B)]}}. \end{aligned} \quad (15.19b)$$

Equation (15.19a) or (15.19b) is a conditional probability so that this model is called conditional logistic model. To avoid confusion, the logistic

Table 15.9 Conditional logistic analysis for screening the risk factors.

Variable	$\hat{\beta}$	$SE(\hat{\beta})$	Wald χ^2	P-value	OR
X_1	0.9445	0.2935	10.3549	0.0013	2.572
X_2	0.8820	0.3242	7.4037	0.0065	2.416

model introduced in the last section is called non-conditional logistic model. One can still use the software to get the maximal likelihood estimates $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$.

In the right-hand side of Eq. (15.19a) or (15.19b), the constant β_0 has been canceled from both of the numerator and denominator so that we are not able to get the estimate $\hat{\beta}_0$ and hence fail to get regression equations (15.18a)–(15.18d). And in fact, (15.19a) or (15.19b) is a conditional probability, which itself is not helpful in predicting the risk of suffering from disease. However, the estimates of coefficients $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ and accordingly $e^{\hat{\beta}_1}, e^{\hat{\beta}_2}, \dots, e^{\hat{\beta}_k}$ have the same meaning as before and can still be used to describe the risk of variables.

For example, a conditional logistic regression was used to screen the risk factors related with stomach cancer at the level of $\alpha = 0.05$, two factors were selected into the model by stepwise method. The results are listed in Table 15.9.

Among three variables, X_3 has not been selected into the model because $P > 0.05$. The results suggest that the dietary habit (X_1) and salted food (X_2) have a relative close association with the occurrence of stomach cancer. Both coefficients of X_1 and X_2 are positive, i.e. $OR > 1$, and of statistical significance, suggesting that they might increase the risk of suffering from stomach cancer. However, the mental status (X_3) is not significantly associated with the risk of stomach cancer.

Example 15.5 is a 1:1 matched case-control design. In principle, the conditional logistic model can be extended to 1: m matched design or other stratified design. However, the calculation will dramatically increase when m increases.

15.3 Multinomial Logistic Regression Model

The logistic regression model introduced in Sec. 15.1 is for a binary dependent variable. However, the variables with more possible outcomes are

frequently faced in medical practice. They may be an ordinal variable such as the outcome of clinical treatment may be described as effective, improvement and not effective; or they may simply be a categorical variable such as pathohistological types of lung cancer are described as squamous-cell carcinoma, adanocarcinoma, small cell carcinoma and alveolar cell carcinoma. In this section, the logistic regression models for these two kinds of outcome variables will be introduced.

15.3.1 Ordinal multinomial logistic model

Example 15.6 In a study, the effects of two medications to certain disease are compared. The outcome of treatment is categorized as effective, improvement and not effective. The data are listed in Table 15.10 stratified by gender. An ordinal multinomial logistic model has been used to analyze the relationship between the outcome of treatment and gender, medication.

There are many types of logistic models for ordinal multinomial variables and most popular one is the cumulative logistic regression model. Assume that there are c possible categories of Y , then $c - 1$ logistic functions are used to describe the relationship between the outcomes and the explanatory variables:

$$\begin{aligned}\text{logit}[P(Y \leq 1)] &= \beta_0^{(1)} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \\ \text{logit}[P(Y \leq 2)] &= \beta_0^{(2)} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \\ &\vdots \\ \text{logit}[P(Y \leq c - 1)] &= \beta_0^{(c-1)} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k. \quad (15.20)\end{aligned}$$

Table 15.10 Data on the impact of gender and medication on certain disease.

Gender	Medication	Effect(Y)			Total
		Effect ($Y = 1$)	Improve ($Y = 2$)	No effect ($Y = 3$)	
Woman ($X_1 = 1$)	New ($X_2 = 1$)	16	5	6	27
	Old ($X_2 = 0$)	6	7	19	32
Man ($X_1 = 0$)	New ($X_2 = 1$)	5	2	7	14
	Old ($X_2 = 0$)	1	0	10	11

Table 15.11 Results of cumulative logistic regression analysis for data in Table 15.10.

Variable	Parameters		Standard error	Wald χ^2	P -value
Intercept	$\hat{\beta}_0^{(1)}$	-2.6672	0.5997	19.7809	0.0001
Intercept	$\hat{\beta}_0^{(2)}$	-1.8128	0.5566	10.6072	0.0011
X_1	$\hat{\beta}_1$	1.3187	0.5292	6.2102	0.0127
X_2	$\hat{\beta}_2$	1.7973	0.4728	14.4493	0.0001

Looking at the right-hand side of (15.20), one will find that among the $c - 1$ equations, the only difference is the intercept term. The statistical software can help to get the maximum likelihood estimates of the commonly shared parameters $\beta_1, \beta_2, \dots, \beta_k$ and the specific parameters $\beta_0^{(1)}, \beta_0^{(2)}, \dots, \beta_0^{(c-1)}$ and related hypothesis tests.

Solution The model has been used for the analysis of the data in Table 15.10. The results are listed in Table 15.11.

The results show that the coefficients of gender and medication are bigger than 0 with small *P* values, i.e. women and new medication have a better outcome of treatment. The results may summarized as two logistic equations:

$$\text{logit}[P(Y \leq 1)] = -2.6672 + 1.3187X_1 + 1.7973X_2,$$

$$\text{logit}[P(Y \leq 2)] = -1.8128 + 1.3187X_1 + 1.7973X_2.$$

These two equations can be applied to predict the prognosis of a patient. For example, suppose a female patient receives a new medication, we have

$$P(Y \leq 1) = e^{(-2.6672+1.3187+1.7973)} = 0.61,$$

$$P(Y \leq 2) = e^{(-1.8128+1.3187+1.7973)} = 0.79.$$

It means the probability of being effective is 0.61, improved is 0.18 and not effective is 0.21 so that the physician may predict her prognosis is not too bad.

15.3.2 Multinomial logistic model

Example 15.7 To study the relationship between the histological types of malignant tumor and cell differentiation and cell staining, a set of data

Table 15.12 Cell differentiation, cell staining and histological types of malignant tumor.

Differentiation (X_1)	Staining (X_2)	Histological types of malignant tumor (Y)		
		Squamous ($Y = 1$)	Adeno ($Y = 2$)	Alveolar cell ($Y = 3$)
I ($X_1 = 1$)	+ ($X_2 = 1$)	10	17	26
	- ($X_2 = 2$)	5	12	50
II ($X_1 = 2$)	+ ($X_2 = 1$)	21	17	26
	- ($X_2 = 2$)	16	12	36
III ($X_1 = 3$)	+ ($X_2 = 1$)	15	15	16
	- ($X_2 = 2$)	12	12	20

was collected (Table 15.12). Apply the multinomial logistic regression to analyze the data.

Here, the histological types of malignant tumor have three categories: squamous, adenous and alveolar cell, denoted by $Y = 1, 2, 3$ respectively. Since the three categorie, Y is not an ordinal variable. A multinomial logistic model can be used to analyze this kind of data.

Assume that there are c possible categories of Y , then Eq. (15.3a) for traditional logistic regression should be extended to $c - 1$ logistic functions in order to describe the relationship between the $c - 1$ possible outcomes and the explanatory variables,

$$\begin{aligned}
 \ln \left[\frac{P(Y = 1)}{P(Y = c)} \right] &= \beta_0^{(1)} + \beta_1^{(1)} X_1 + \cdots + \beta_k^{(1)} X_k, \\
 \ln \left[\frac{P(Y = 2)}{P(Y = c)} \right] &= \beta_0^{(2)} + \beta_1^{(2)} X_1 + \cdots + \beta_k^{(2)} X_k, \\
 &\dots \\
 \ln \left[\frac{P(Y = c - 1)}{P(Y = c)} \right] &= \beta_0^{(c-1)} + \beta_1^{(c-1)} X_1 + \cdots + \beta_k^{(c-1)} X_k. \quad (15.21a)
 \end{aligned}$$

Accordingly, Eqs. (16.3c) and (16.3d) are extended as

$$\begin{aligned}
 P(Y = 1) \\
 &= \frac{e^{(\beta_0^{(1)} + \beta_1^{(1)} X_1 + \cdots + \beta_p^{(1)} X_k)}}{1 + e^{(\beta_0^{(1)} + \beta_1^{(1)} X_1 + \cdots + \beta_p^{(1)} X_k)} + \cdots + e^{(\beta_0^{(c-1)} + \beta_1^{(c-1)} X_1 + \cdots + \beta_p^{(c-1)} X_k)}}.
 \end{aligned}$$

$$\begin{aligned}
P(Y = 2) &= \frac{e^{(\beta_0^{(2)} + \beta_1^{(2)} X_1 + \dots + \beta_p^{(2)} X_k)}}{1 + e^{(\beta_0^{(1)} + \beta_1^{(1)} X_1 + \dots + \beta_p^{(1)} X_k)} + \dots + e^{(\beta_0^{(c-1)} + \beta_1^{(c-1)} X_1 + \dots + \beta_p^{(c-1)} X_k)}} \\
&\vdots \\
P(Y = c - 1) &= \frac{e^{(\beta_0^{(c-1)} + \beta_1^{(c-1)} X_1 + \dots + \beta_p^{(c-1)} X_k)}}{1 + e^{(\beta_0^{(1)} + \beta_1^{(1)} X_1 + \dots + \beta_p^{(1)} X_k)} + \dots + e^{(\beta_0^{(c-1)} + \beta_1^{(c-1)} X_1 + \dots + \beta_p^{(c-1)} X_k)}}, \\
P(Y = c) &= \frac{1}{1 + e^{(\beta_0^{(1)} + \beta_1^{(1)} X_1 + \dots + \beta_p^{(1)} X_k)} + \dots + e^{(\beta_0^{(c-1)} + \beta_1^{(c-1)} X_1 + \dots + \beta_p^{(c-1)} X_k)}}.
\end{aligned} \tag{15.21b}$$

Same as those in (15.20), we have $c - 1$ equations in (15.21a); but instead of the logarithm of odds for $P(Y \leq m)$, the left-hand side here are the logarithm of relative risks, which are the ratios between $P(Y = m)$ and $P(Y = c)$, $m = 1, 2, \dots, c - 1$. The right-hand side of (15.21a) shows that the number of parameters in (15.21a) is much more than that in (15.20), besides the intercepts $\beta_0^{(1)}, \beta_0^{(2)}, \dots, \beta_0^{(c-1)}$, the regression coefficients are also different for different equations. We have so many parameters to estimate in this model just because that the different categories of Y may be quite different in nature. However, as long as the sample size is big enough, the statistical software can still help us to get the maximum likelihood estimates of the unknown parameters.

Solution The above model has been used for the analysis of the data in Table 15.12. The results are listed in Table 15.13.

By (15.21a), two equations are used to describe the relationship between squamous cancer and alveolar cell cancer and between adenocarcinoma and alveolar cell cancer. They are

$$\begin{aligned}
\ln \left[\frac{P(Y = 1)}{P(Y = 3)} \right] &= \beta_0^{(1)} + \beta_1^{(1)} X_1 + \beta_2^{(1)} X_2, \\
\ln \left[\frac{P(Y = 2)}{P(Y = 3)} \right] &= \beta_0^{(2)} + \beta_1^{(2)} X_1 + \beta_2^{(2)} X_2.
\end{aligned}$$

Table 15.13 Multinomial logistic regression analysis for Example 15.7.

Variable	Parameter	Standard error	Wald χ^2	P-value	$RR = \exp(\beta)$	
Intercept	$\hat{\beta}_0^{(1)}$	-0.9826	0.5707	2.96	0.0851	
	$\hat{\beta}_0^{(2)}$	-0.3461	0.5413	0.41	0.5226	
X_1	$\hat{\beta}_1^{(1)}$	0.6281	0.1799	12.19	0.0005	1.874
	$\hat{\beta}_1^{(2)}$	0.3454	0.1728	4.00	0.0456	1.413
X_2	$\hat{\beta}_2^{(1)}$	-0.6494	0.2833	5.26	0.0219	0.522
	$\hat{\beta}_2^{(2)}$	-0.6352	0.2725	5.43	0.0197	0.530

The maximum likelihood estimates of the parameters are listed in Table 15.13.

For the test of goodness of fit we have $G = 7.39$, P -value is 0.2864 so that the multinomial logistic model fits these data. The last column of Table 15.13 shows that one grade increase of cell differentiation will cause the relative risk of squamous cancer and alveolar cell cancer becoming 1.874 times as much as that before increasing, and the relative risk of adenocarcinoma and alveolar cell cancer becoming 1.413 times as much as that before increasing; while negative staining may cause both of the relative risks becoming 0.522 and 0.530 times as much as that for positive staining. In other words, it tends to be squamous cancer or adenocarcinoma if the differentiation is higher; and it tends to be alveolar cell cancer if the staining is negative.

By (15.21a), the results may be summarized as

$$\ln \left[\frac{P(Y = 1)}{P(Y = 3)} \right] = -0.9826 + 0.6281X_1 - 0.6494X_2,$$

$$\ln \left[\frac{P(Y = 2)}{P(Y = 3)} \right] = -0.3461 + 0.3454X_1 - 0.6352X_2$$

or by (15.21b)

$$P(Y = 3)$$

$$= \frac{1}{1 + e^{(-0.9826+0.6281X_1-0.6494X_2)} + e^{(-0.3461+0.3454X_1+-0.6352X_2)}},$$

$$\Pr(Y = 2)$$

$$= \frac{e^{(-0.3461+0.3454X_1+-0.6352X_2)}}{1 + e^{(-0.9826+0.6281X_1-0.6494X_2)} + e^{(-0.3461+0.3454X_1+-0.6352X_2)}},$$

$$P(Y = 3)$$

$$= \frac{1}{1 + e^{(-0.9826+0.6281X_1-0.6494X_2)} + e^{(-0.3461+0.3454X_1+-0.6352X_2)}}.$$

15.4 Two-Level Logistic Mixed Effects Regression Model

Both in epidemiological study and clinical research, we often encounter with data which has hierarchical structure. The general example is stratified random sampling by area and individuals, in such a sample we have two levels area and individual. In fact, lots of experimental design can produce hierarchical structure data, such as the centers and subjects in multicenter clinical trials, the nest and animals in toxicology study. The characteristics of hierarchical structure is that the effects between groups is larger however the effects within groups is smaller, in other words, the data have certain cluster effect. When we get a data with hierarchical structure, the traditional logistic regression model is not suitable any more for the data do not meet the assumption of independence. If we ignore the hierarchical structure of data and use traditional logistic regression model, there may be bias and get a wrong conclusion.

Now we will use a medical service survey data which comes from the book *The Multilevel Statistical Model in Medical and Public Health* edited by Yang Min and Li Xiao-Song to illustrate the two-level logistic regression model.

Example 15.8 For screening the factors which affect the health services demand in poor rural areas, a stratified random sampling was used, first we select 832 families as our object families then we randomly choose 2369 residents aged 15 years or older who come from the selected families as our participants. The dependent variable is “uncomfort”, which refers to suffering diseases within two weeks; other variables in the data are dealt as independent variables. The variables and their values are listed in Table 15.14.

Table 15.14 Variable names and its assignment.

Variable	Labels	Assignment
discomfort	Diseases within two weeks	0=NO 1=YES
chronic	Chronic diseases	0=NO 1=YES
gender	Gender	0=Male 1=Female
agegroup	Age(year)	0=15 1=45 2=65
marriage	Marriage	0=Single 1=Married 2=Divorce 3=Lose partner
edu	Education	0= Illiteracy or semi-illiteracy 1=Primary 2=Middle school or higher
occup	Occupation	0=Workers 1=farmers 2=Student 3=Retired 4=Unemployed
smoke	Smoke	0=NO 1=YES
drink	Drink	0=No or little 1=Frequent
water	Drinking water type	0=Tap water 1=Non Tap water
geography	Geography type	0=Mountain area 1=Non mountain area
id	Individual level	Level 1
family	Family level	Level 2

As age, marital status, education and occupation are polytomous variables, we assign them by dummy variables with value zero as reference for contrast:

Age 1: "45", Age 2: "65"

Marriage1: married, Marriage2: divorce, Marriage3: lose partner

Edu1: primary school, Edu2: middle school or higher

Occup1: farmer, Occup2: student, Occup3: retire, Occup4: unemploy

If we ignore the fact that the individuals come from different families then (15.3a) is an appropriate model of this example. However, when the variation among families is larger than the variation within families, we cannot view the 2369 residents as coming from the same population anymore. When we study the sampling error, because every family contains a number of individuals, we call the variation among individuals within family with level 1 variation, and call the variation among families with level 2 variation. If we take the variation among families into account, the logistic regression model should be extended to (15.22)

$$\text{logit}(P) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u_0, \quad (15.22)$$

where u_0 refers to the family effect, which contributes to $\text{logit}(P)$. We may assume u_0 follow a normal distribution $u_0 \sim N(0, \sigma_{u_0}^2)$.

A statistical test is needed for

$$H_0 : \sigma_{u_0}^2 = 0 \quad H_1 : \sigma_{u_0}^2 \neq 0.$$

If it cannot be rejected, then the effect at level 2 can be ignored, otherwise not.

For comparing which model is more suitable to Example 15.8, we construct an empty model which contains no explanatory variables to estimate if the family effect is ignorable.

After running the SAS Program 15.5 at the end of this chapter, we get the results in Table 15.15, from which we conclude that the family effect in the empty model is not zero ($P < 0.001$).

Then we put the explanatory variables into the empty model step by step to explore which factors affect the health services demand in poor rural areas. The results are showed in Table 15.16.

Table 15.15 The empty two-level model for diseases within two weeks.

Variables	Parameters	Values	SE	<i>t</i>	<i>P</i>
Fixed effect					
Intercept	β_0	-1.272	0.072	-17.59	<0.001
Random effect					
Variance at level 2	$\sigma_{u_0}^2$	1.104	0.240	4.60	<0.001

Table 15.16 The random effect model for diseases within two weeks.

Variables	Parameters	Values	SE	<i>t</i>	<i>P</i>
Fixed effect					
Intercept	β_0	-2.659	0.163	-16.34	<0.001
Age (Year) 45–		0.612	0.149	4.11	<0.001
65–		0.846	0.197	4.29	<0.001
Gender		0.400	0.132	3.04	0.002
Chronic diseases		2.843	0.188	15.11	<0.001
Drink		0.506	0.239	2.12	0.034
Random effect					
Variance at level 2	$\sigma_{u_0}^2$	1.258	0.347	3.630	0.003
Random coefficient (Drink)		0.443	1.135	0.390	0.696

From Table 15.16 we know that, the factors affect the health services demand in poor rural areas is age, gender, chronic diseases and drink. The health services demand rate for “45” and “65” groups is higher than group “15”; women higher than man; have chronic diseases within half year higher than non-chronic diseases within half year; drinking higher than non-drinking.

Since the p value is 0.003, the test for $H_0 : \sigma_{u_0}^2 = 0$ can be rejected, which reminds us that the health services demand varied sharply at family level.

15.5 Application of Logistic Regression

Recently logistic regression model has been widely used in all fields of medical research, including data analysis of cohort study and case-control study in epidemiology and etiology, discriminant model in clinical diagnostic test and evaluation of treatment effect in clinical study. The applications of logistic regression model may be summarized as the following three aspects.

15.5.1 Screening of disease-related risk factors

As the development of medicine, it is emphasized by different approaches to explore the causes of disease. As discussed before, logistic regression model has its advantage in multiple factor analysis of disease etiological study. It is suitable to screen disease-related risk factors from many potential factors, as well as to analyze the interaction between different risk factors. Example 15.3 is a typical case of risk factor screening.

15.5.2 Adjustment of confounding factors

In the studies of clinical medicine and epidemiology, there are confounding impacts of some non-study-interest factors on study factors frequently. For example, age, gender, status of disease and stage of disease, etc. may bias the evaluation of treatment effect and age, occupation and income, etc. may affect the analysis for the relationship between disease occurrence and living habits. There are two methods to deal with the influence of confounding effect. One is to control in study design, such as applying stratified sampling, matched design or randomized block design to balance the confounding factors between experiment group and control group. Another

is to control in statistical analysis, like classical method of Mantel–Haenszel stratified analysis. However, Mantel–Haenszel method can only deal with data of $2 \times 2 \times K$ table so that it is not proper when study factors have multiple levels or there are many confounding factors. It is very convenient to control the confounding factors and to estimate the odds ratio and confidence interval by logistic regression model. Especially when there are many confounding factors, logistic regression model may sufficiently use the information within data. Examples 15.2 and 15.3 are typical cases of confounding control.

15.5.3 Prediction and discrimination

Same as other regression analysis, logistic regression model may be used for prediction. Logistic regression model is a probability model and the probability of an event occurring under certain condition can be calculated as in Examples 15.1 and 15.3. More than that, it can be applied in differential diagnosis in clinical medicine.

15.5.4 Attention should be paid

In the logistic analysis, we should pay attention to the following situation:

15.5.4.1

Sometimes a logistic regression may include several or more than 20 independent variables. As the increase of number of independent variables, the number of cross categorized levels between variables will increase rapidly so that an enough sample size is needed to ensure the stability of parameter estimation. Otherwise, the estimated value of parameter might be very large or very small and difficult to explain.

15.5.4.2

The screening of risk factors should not totally depend on computer and a fixed significance level. Clinical or epidemiological meanings and explanation of results are more important. The users may choose some important risk factors into the model by themselves based on their clinical or epidemiological knowledge.

15.5.4.3

The independent variables in logistic regression model may be categorical variable, ordinal variable and continuous variable. If the independent variable is a multi-level categorical variable, we may use several dummy variables to replace it and make the final results more easy to explain. The regression coefficient of continuous variable is also difficult to explain so that it is often transformed into an Ordinal variable.

15.5.4.4

The intercept of the model is meaningless in many cases and do not need to explain its meaning or test its significance. Only in cohort study and cross-sectional study, the frequency of the event in sample is close to the probability in general population, the intercept of model is meaningful. In conditional logistic regression model, the intercept has been canceled so that it cannot be used for prediction.

15.6 Computerized Experiments

Experiment 15.1 Nonconditional logistic regression model The SAS Program 15.1 is for the logistic regression analysis of Example 15.2. Lines 04 to 15 are the data: the first column is for the values of dependent variable *Y*; the second column is for the values of independent variable SVEG (salted vegetable intake); the third to seventh columns are for the

Program 15.1 Non-conditional logistic regression model.

Line	Program	Line	Program
01	DATA A;	12	0 1 0 1 0 0 0 29 0 0 0 1 0 0 0 1 38
02	INPUT Y SVEG A1 A2 A3 A4 A5	13	0 1 0 0 1 0 0 27 0 0 0 0 1 0 0 1 38
	COUNT@@;	14	0 1 0 0 0 1 0 18 0 0 0 0 0 1 0 88
03	CARDS;	15	0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 31
04	1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0	16	
05	1 1 1 0 0 0 0 4 1 0 1 0 0 0 0 5	17	PROC LOGISTIC
06	1 1 0 1 0 0 0 25 1 0 0 1 0 0 0 21		DESCENDING;
07	1 1 0 0 1 0 0 42 1 0 0 0 1 0 0 34	18	FREQ COUNT;
08	1 1 0 0 0 1 0 19 1 0 0 0 0 1 0 36	19	MODEL Y=A1 A2 A3 A4 A5
09	1 1 0 0 0 0 1 5 1 0 0 0 0 0 1 8		SVEG;
10	0 1 0 0 0 0 0 9 0 0 0 0 0 0 0 106	20	RUN;
11	0 1 1 0 0 0 0 26 0 0 1 0 0 0 0 164		

five age dummy variables; column 8 is for the frequency. Line 17 runs the procedure of logistic regression. Line 18 defines the frequency weighting variable and line 19 defines the model structure.

Experiment 15.2 Conditional logistic regression model The SAS Program 15.2 is used for the conditional logistic regression analysis of Example 15.4.

Program 15.2 Conditional logistic regression model.

Line	Program
01	DATA A;
02	INPUT NO Y X1 X2 X3@@;
03	CARDS;
04	1 0 2 4 0 1 1 3 1 0 2 6 0 2 2 0 2 6 1 1 1 0
05	2 0 3 2 1 2 1 0 1 0 2 7 0 2 0 1 2 7 1 0 2 1
06	3 0 3 0 0 3 1 2 0 1 2 8 0 1 1 1 2 8 1 3 0 1
07	4 0 3 0 0 4 1 2 0 1 2 9 0 2 0 1 2 9 1 4 0 0
08	5 0 3 0 1 5 1 0 0 0 3 0 0 3 1 0 3 0 1 0 2 1
09	6 0 2 2 0 6 1 0 1 0 3 1 0 1 0 1 3 1 1 0 0 0
10	7 0 3 1 0 7 1 2 1 0 3 2 0 4 2 1 3 2 1 1 0 1
11	8 0 3 0 0 8 1 2 0 0 3 3 0 4 0 1 3 3 1 2 0 1
12	9 0 2 2 0 9 1 1 0 1 3 4 0 2 0 1 3 4 1 0 0 1
13	10 0 1 0 0 10 1 2 0 0 3 5 0 1 2 0 3 5 1 2 0 1
14	11 0 3 0 0 11 1 0 1 1 3 6 0 2 0 0 3 6 1 2 0 1
15	12 0 3 4 0 12 1 3 2 0 3 7 0 0 1 1 3 7 1 1 1 0
16	13 0 1 1 1 13 1 2 0 0 3 8 0 0 0 1 3 8 1 4 0 0
17	14 0 2 2 1 14 1 0 2 1 3 9 0 3 0 1 3 9 1 0 1 0
18	15 0 2 3 0 15 1 2 0 0 4 0 0 2 0 1 4 0 1 3 0 1
19	16 0 2 4 1 16 1 0 0 1 4 1 0 2 0 0 4 1 1 1 0 1
20	17 0 1 1 0 17 1 0 1 1 4 2 0 3 0 1 4 2 1 0 0 1
21	18 0 1 3 1 18 1 0 0 1 4 3 0 2 1 1 4 3 1 0 0 0
22	19 0 3 4 1 19 1 2 0 0 4 4 0 2 0 1 4 4 1 1 0 0
23	20 0 0 2 0 20 1 0 0 0 4 5 0 1 1 1 4 5 1 0 0 1
24	21 0 3 2 1 21 1 3 1 0 4 6 0 0 1 1 4 6 1 0 0 0
25	22 0 1 0 0 22 1 2 0 1 4 7 0 2 1 0 4 7 1 0 0 0
26	23 0 3 0 0 23 1 2 2 0 4 8 0 2 0 1 4 8 1 1 1 0
27	24 0 1 1 1 24 1 0 1 1 4 9 0 1 2 1 4 9 1 0 0 1
28	25 0 1 2 0 25 1 2 0 0 5 0 0 2 0 1 5 0 1 0 3 1
29	;
30	PROC PHREG;
31	MODEL Y=X1 X2 X3/SELECTION=STEPWISE SLENTTRY=0.05;
32	STRATA NO;
33	RUN;

Program 15.3 Ordinal multinomial logistic regression model.

Line	Program
01	DATA A;
02	INPUT Y X1 X2 COUNT@@;
03	CARDS;
04	1 1 1 16 1 0 1 5
05	2 1 1 5 2 0 1 2
06	3 1 1 6 3 0 1 7
07	1 1 0 6 1 0 0 1
08	2 1 0 7 2 0 0 0
09	3 1 0 19 3 0 0 10
10	;
11	PROC LOGISTIC;
12	FREQ COUNT;
13	MODEL Y= X1 X2/SCALE=NONE AGGREGATE;
14	RUN;

Experiment 15.3 Ordinal multinomial logistic regression model model

Program 15.3 is used for ordinal multinomial logistic regression analysis of Example 15.5. In Program 15.2, NO is the number of pairs. Procedure PHREG is used for analysis of proportional hazards model and statement STRATA tells the computer that NO is a variable of strata. Computer develops conditional probability model according to the value of stratum variable. In the options of the statement MODEL, SELECTION defines the method for variable screening (here is stepwise method) and SLENTY defines the significance level of variable screening (here is 0.05).

Experiment 15.4 Multinomial logistic regression model Program 15.4 is used for multinomial logistic regression analysis of Example 15.6.

Experiment 15.5 Two-level Logistic regression model Taking Example 15.8 as an example, master the computer manipulation and SAS procedures for two-level logistic regression model (run SAS procedure GLIMMIX needed SAS9.2 or higher version).

In Program 15.5, 01–10 lines were Proc NLMIXED procedures for constructing two-level empty model and estimate if there is any high level effects also called the statistical test for $\sigma_{u_0}^2$ and ICC. Lines 12–17 were Proc GLIMMIX procedures for constructing the ultimate two-level random effect model and proving the primary parameter estimate. Lines 19–34 were Proc NLMIXED procedures for constructing the ultimate regression model.

Program 15.4 SAS program of multinomial logistic regression model.

Line	Program
01	DATA A;
02	DO X1=1 TO 3;
03	DO X2=1 TO 2;
04	DO Y=1 TO 3;
05	INPUT COUNT@@;
06	OUTPUT;
07	END;END;END;
08	CARDS;
09	10 17 26 5 12 50 21 17 26 16 12 26 15 15 16 12 12 20
10	;
11	PROC CATMOD ORDER=DATA;
12	DIRECT X1 X2;
13	WEIGHT COUNT;
14	MODEL Y=X1 X2;
15	RUN;

Program 15.5 Two-level Logistic regression model.

Line	Program
01	title "Two level Logistic regression empty model";
02	Proc NLMIXED data=health;
03	PARMS B0=0 V_u0=1 ;
04	logodds=B0+u0j;
05	odds=exp(logodds);
06	P= odds/(1+odds);
07	model uncomfor~binary(P);
08	random u0j ~normal(0,V_u0) subject=family;
09	estimate 'ICC' V_u0/(V_u0+3.289);
10	run;
11	title "Two level Logistic regression model initial Coefficients";
12	Proc GLIMMIX data=health method= RSPL;
13	class family;
14	Model uncomfor (event='1') =AGEGROU1 AGEGROU2 gender chronic drink marriage1 marriage2 marriage3 edu1 edu2 OCC1 OCC2 OCC3 OCC4/ s dist = binary link =logit DDFM=BW;
15	random int / subject=family s cl;
16	nloptions tech= NRRIDG;
17	Run;
18	title "Two level Logistic regression random effect model"

(Continued)

Program 15.5 (Continued)

Line	Program
19	Proc NLMIXED;
20	PARMS B0=-3 B1=0.5 B2=0.9 B3=0.3 B4=2 B5=0.5 t11=0.7 t22=0.3 t12=0;
21	Z=B0+B1*AGEGROU1+B2*AGEGROU2+B3*gender+B4*chronic+ B5*drink+(u0j+u1j*drink);
22	if(uncomfor=1) then P=1/(1+EXP(-Z));
23	ELSE P=1-(1/(1+EXP(-Z)));
24	LL=LOG(P);
25	V_u0=t11*t11;
26	cov_u01=t11*t12;
27	v_u1=t12*t12+t22*t22;
28	model uncomfor~general(LL);
29	random u0j u1j~normal([0,0],[V_u0,cov_u01,v_u1]) subject=family;
30	estimate 'ICC' V_u0/(V_u0+3.289);
31	estimate 'VAR(u0)' V_u0;
32	estimate 'cov(u0,u1)'cov_u01;
33	estimate 'VAR(u1)'v_u1;
34	run;

Table 15.17 The values of age in Example 15.3.

Age	25—	35—	45—	55—	65—	75+
AGE	1	2	3	4	5	6
A1	0	1	0	0	0	0
A2	0	0	1	0	0	0
A3	0	0	0	1	0	0
A4	0	0	0	0	1	0
A5	0	0	0	0	0	1

15.7 Practice and Experiments

1. In Example 15.3, the values of age have been given by two different ways as show in Table 15.17.

If AGE is adopted into the logistic regression model, what is the meaning of the regression coefficient? If A1, A2, A3, A4, A5 are adopted into the logistic regression model, what are the meaning of the regression coefficients? What is the difference and relation between these two sets of regression coefficients for age? Which way is better? Why?

2. When there are data from a 1:1 pair-designed case-control study, what will influence the results if non-conditional logistic regression is used?
3. In a study on the relationship between disease occurrence and X_1 , X_2 , when X_1 or X_2 is included into the model respectively, both of the two logistic regressions are of significance but when both of them are included into the model simultaneously, the logistic regression is not significant. Why?
4. Try an experiment by yourself with the data of Example 15.2: calculate the difference of exposure values of case minus that of control within each pair; regard each pair as a "case" with the difference as the value of explanatory variable X and with $Y = 1$ as the "outcome" of the "case". Perform a non-conditional logistic regression analysis on such a "dataset". Are the results different from that obtained by a conditional logistic regression?
5. In a cohort study of 1000 workers in a factory, 194 workers have a white blood cell reduction among 900 workers exposed to the risk factor benzene and only 21 workers have the same symptom among 100 workers not exposed to the risk factor. In another cohort study of 1000 workers in another factory, 6 have symptom among 100 exposed workers and 29 among 900 non-exposed workers. The data are pooled in Table 15.18.

Try to use the classical M-H method to estimate the odds ratio of the benzene exposure and to do a chi-square test; and then use logistic regression to estimate the odds ratio of the benzene exposure and to do a score test. Compare the results of two different analyses and discuss how to deal with the difference of results when analyze these data separately and pooling together.

Table 15.18 Pooled data of the two factories.

	E+	E-	Total
D+	200	50	250
D-	800	950	1750
Total	1000	1000	2000

(1st edn. Qing Liu, Jiqian Fang; 2nd edn. Jinxin Zhang, Weidong Li, Jiqian Fang)



Chapter 16

Cluster Analysis

Cluster analysis is a kind of methods to group individuals or variables with similar properties. It is used widely in biology and medical field for classifications. For example, according to the morphology, size, and numbers of bones, cluster analysis can group the evolution process into several periods; health administrator classifies the hospitals into different grades according to their diagnostic accuracy, treatment capacity etc. To restore ears, doctors classify normal ears into different categories according to the length, width and some distances, etc., and then find several standardized ears for patients with impaired ones.

16.1 The Meaning of Clustering

16.1.1 *Comparison of cluster analysis and discriminant analysis*

Cluster analysis differs from the discriminant analysis in the following sense. In a discriminant analysis, there is a gold standard that defines the belongings of all study subjects. The discriminant rule is built based on the knowledge of true classifications of training dataset (supervised). On the other hand, the number of groups and their labels/characteristics are unknown and there is no known class variable in cluster analysis (unsupervised). In some cases, we call the discriminant analysis *supervised learning* and the cluster analysis *unsupervised learning*. In other cases, we combine two methods in applications: we first use cluster analysis to classify the subjects into several different groups and then use discriminant analysis to find the classification rules for future observations.

Table 16.1 Data structure for cluster analysis.

Individuals	Variables
	$X_1 X_2 \dots X_j \dots X_m$
1	$X_{11} X_{12} \dots X_{1j} \dots X_{1m}$
2	$X_{21} X_{22} \dots X_{2j} \dots X_{2m}$
\vdots	\vdots
i	$X_{i1} X_{i2} \dots X_{ij} \dots X_{im}$
\vdots	\vdots
n	$X_{n1} X_{n2} \dots X_{nj} \dots X_{nm}$
Mean	$\bar{X}_1 \bar{X}_2 \dots \bar{X}_j \dots \bar{X}_m$
Standard Deviation	$S_1 S_2 \dots S_j \dots S_m$

16.1.2 Cluster analysis

The data for cluster analysis should be organized in the form showed in Table 16.1.

There are a few approaches for cluster analysis, which are all based on the same principal of grouping data according to their closeness or similarity. Here are some commonly used methods:

- (1) Hierarchical cluster, used for cluster analysis of small number of individuals or variables.
- (2) Fast cluster, used for large number of individuals or variables.
- (3) Cluster of variables.
- (4) Cluster of ordered individuals, used to group individuals in the order that an individual will group only with individual next to him/her.

16.1.3 Cluster statistics

The kind of statistics that measure the closeness of individuals or variables are called cluster statistics. The most commonly used cluster statistics are the distance and correlation coefficients.

(1) *Distance*: Distance is used to cluster individuals. For any two individuals i and k , their *Euclidean distance* is defined as

$$d_{ik} = \sqrt{(x_{i1} - x_{k1})^2 + (x_{i2} - x_{k2})^2 + \cdots + (x_{im} - x_{km})^2}$$

$$= \left[\sum_{j=1}^m (x_{ij} - x_{kj})^2 \right]^{1/2}.$$

Here, x_{ij} and x_{kj} are the values of the j th variable for the i th and k th individuals, respectively. Sometimes we also use the square of Euclidean distance d_{ik}^2 .

In order to remove the unit of variables, we should standardize the variables before we calculate the distance. A common way is to replace all variables in the unit of standard deviations, i.e., use the transformed variable

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j},$$

where \bar{x}_j and S_j are the sample mean and sample standard deviation for the j th variable. The standardized variable should have sample mean zero and sample standard deviation 1. The distance calculated using the standardized variable will not change by the alteration of the measurement unit.

(2) *Correlation coefficients*: Correlation coefficients are used for variable cluster analysis. For variables X_i and X_k , the correlation coefficient r_{ik} is

$$r_{ik} = \frac{\sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\left[\sum_{j=1}^n (x_{ji} - \bar{x}_i)^2 \right] \left[\sum_{j=1}^n (x_{jk} - \bar{x}_k)^2 \right]}}.$$

16.2 Hierarchical Cluster

Hierarchical cluster is a commonly used method. The basic idea is the following. First, we define the distance between individuals and clusters. At the beginning, each individual is a cluster. We combine the two clusters with the minimum distance into a new cluster and then recalculate the distances between new cluster and all the other clusters. We, again, combine the two clusters with the minimum distance into one cluster and recalculate the

new distances. Each time we repeat this procedure, the cluster number is reduced by 1 until all individuals are in one cluster. The whole process can be displayed by a *tree graph* with branches representing the clusters. Using different definitions of distances one may result in different trees and, thus, different set of clusters. In practice, we can use several distances to generate several trees; then compare these trees and use the background knowledge to select one that is most biologically reasonable.

SAS procedure PROC CLUSTER can be used to perform individual cluster analysis. It has 11 different distances to choose from, and displays the resulting trees.

Suppose we have two individuals from cluster A , as A_1 and A_2 , respectively in Fig. 16.1. We also have three individuals, B_1 , B_2 , and B_3 , from cluster B .

The single linkage uses the smallest Euclidean distance between individuals from cluster A and individuals from cluster B as the distance between the two clusters. In Fig. 16.1 it is the distance between A_2 and B_1 .

The complete linkage uses the largest Euclidean distance between individuals from cluster A and individuals from cluster B as the distance between two clusters. In Fig. 16.1 it is the distance between A_2 and B_3 .

In the centroid method, the distance between two clusters is defined as the (squared) Euclidean distance between their centroids or means. In Fig. 16.1, it is calculated as the (squared) Euclidean distance between the midpoint of A_1 and A_2 and the centroid of B_1 , B_2 , and B_3 .

In average linkage the distance between two clusters is the average squared Euclidean distance between pairs of individuals, one in each cluster. In Fig. 16.1, it is the average of (squared) Euclidean distance

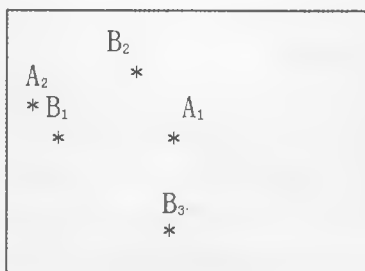


Fig. 16.1 Illustration of several definitions of distance between two clusters.

between A_1 and B_1 , A_1 and B_2 , A_1 and B_3 , A_2 and B_1 , A_2 and B_2 and A_2 and B_3 .

In addition, there are other methods, such as *median method*, *flexible-beta method*, *McQuitty's similarity analysis*, *Ward's minimum-variance method*, *EML method*, *density linkage* and *two-stage density linkage*, etc.

Example 16.1 This study intended to use bacteria total fatty-acid measured by gas chromatogram to analyze and cluster several types of bacteria. 24 bacteria species were collected, including eight types of vibrio jejuni (denoted as J1–J8), three types of vibrio colon (denoted as CC1–CC3), nine types of spirobacteria pylorus denoted as (HP1–HP9), and four other enterobacteria denoted as XX1–XX4. For each bacteria species, 12 fatty acid, denoted as X1–X12 were recorded. The goal was to classify the 24 bacteria species using variables X1–X12.

In this study, the number of individuals was only 24, which was small enough to use hierarchical cluster method. Using 11 distance methods in the SAS PROC CLUSTER, we obtained 11 trees. Most algorithms made three clusters. Among them, the results of average linkage, EML, and Ward's minimum-variance methods had reasonable cluster results, which were similar to the clusters using other microbiologic methods.

Now we use the average linkage method as an example. The resulted tree is given in Fig. 16.2. In a tree, the objects that are clustered, either individuals or variables, are *leaves*; the cluster containing all objects is the *root*; a cluster containing at least two objects but not all of them is a *branch*; the general term for leaves, branches, and roots is *node*. If a cluster A is the union of clusters B and C , then A is the *parent* of B and C , and B and C are *children* of A . A leaf is thus a node with no children, and a root is a node with no parent. If every cluster has at most two children, the tree is a *binary tree*.

Figure 16.2 is a horizontal tree where the name of the individuals is displayed on the vertical axis and the distance between individuals is displayed on the horizontal axis. The root is on the left and leaves on the right. Individuals are grouped from right to left. Every observed individual is a leaf in the most right column. If a sample has not yet been combined with others into a branch, we denote it as "O". Once it is combined with others into a branch (i.e. a cluster), we use the notation "X". Two clusters are separated by an empty row. As you look up from the right of the diagram, individuals and



Fig. 16.2 Cluster tree using average linkage method.

clusters are progressively joined until a single, all-encompassing cluster is formed on the left (or root) of the diagram. Clusters exist at each level of the diagram. For example, the unbroken line of Xs at the top three rows of the 0.3 level indicates that the two bacteria HP1 and HP6 have formed a cluster. The next cluster is between HP1, HP6, and HP5, at distance 0.85, etc. This tree plot is different from the tree diagrams presented in the literature. With the high resolution graphics option in SAS PROC TREE or based on the tree plot in Fig. 16.2, we can construct a tree using the horizontal and vertical lines to clearly display the process, like Fig. 16.3 in our example.

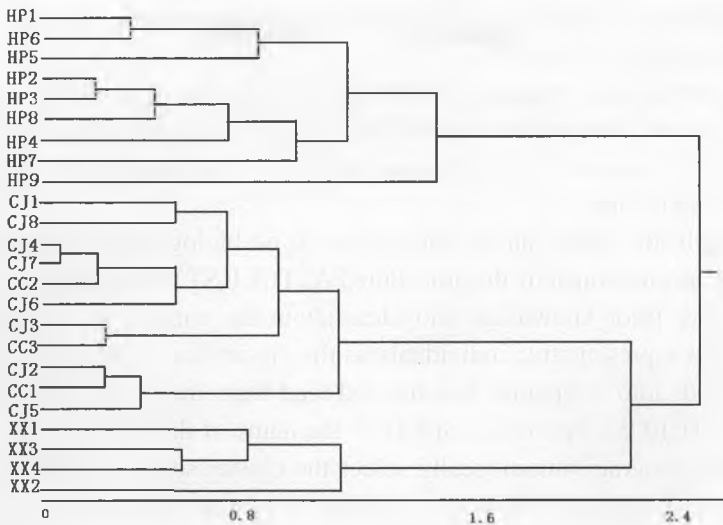


Fig. 16.3 Tree diagram of average linkage method.

In Fig. 16.2, we can see that when the distance equals 2.2, there are three clusters with 9 HP bacteria, 8 CJ and 3 CC bacteria, and 4 XX bacteria. This cluster classification can be confirmed by other methods and therefore can be accepted with reasonable confidence. Using other methods can also generate three clusters with similar results.

16.3 Fast Cluster

When sample size is very large, the computational burden of the hierarchical cluster method is heavy and the tree diagram can be too complicated to understand. Therefore, we can use SAS PROC FASTCLUS to save computation time for large sample cluster. Because FASTCLUS procedure does not automatically standardize variables, if the variables used to cluster individuals are all measured in the same unit, it should be fine; otherwise, if the variables are not measured in the same unit, they should be standardized before their uses in this procedure. The FASTCLUS procedure uses Anderberg's nearest centroid method. It first selects a few starting cluster seeds and uses them as the centroids for the clusters; then it calculates all the distances between any individual and these centroids and clusters

the individual to the cluster corresponding to the nearest centroid. This generates temporary clusters.

Based on these temporary clusters, the algorithm calculates the new centroids and, then, reclassifies the individuals based on updated centroids. The algorithm iterates this procedure until there is no change in clustering of the individuals.

The cluster seeds can be selected based on biological knowledge and clinical information or by the procedure FASTCLUST without human input. If we have prior knowledge and ideas about the number of clusters, we can select representative individuals as the cluster seeds. We can compile these seeds into a separate data file and read them into the program using the FASTCLUST option of "SEED=" the name of data file. If we decide to let the program automatically select the cluster seeds, we can use the option "MAXCLUSTERS=" to a reasonable number of maximum clusters. Different maximum initial numbers can give different clustering results. Therefore, the selection of initial maximum number should accord to all information available to your specific problem.

Example 16.2 The rehabilitation clinic of a hospital measured 300 normal ears in order to develop some standards of normal ears. These standards would be used as the references for ear repairing of ear trauma patients. For these 300 normal ears, they measured ear length (EC), ear width (EK), distance between helix and anthelix (EZ), ear shape (EX), and length of earlobe (ECX). In addition to these direct measures, they also used derived parameters: the ear index ($EI = (EK/EC) \times 100\%$) and Anthelix index ($AI = (EZ/EW) \times 100\%$). The goal was to find representative clusters of normal ears.

Solution As 300 ears were too large for hierarchical cluster method, we should use the fast cluster method. After using the SAS procedure PROC STANDARD to standardize the measures, the PROC FASTCLUST was used to perform a cluster analysis. The initial cluster seeds were selected by the procedure. According to a survey of clinicians, the number of clusters should not be more than four. The clusters resulted from this analysis were given in Table 16.2. In this table, the means of four clusters were listed. They represented the average directions of each cluster with regard to the overall means in the unit of standard deviation. The original scales could be calculated backwards by multiplying the standard deviation and then

Table 16.2 The frequencies and means of the four clusters in Example 16.2.

Cluster	Frequency	EC	EK	EZ	EI	AI
1	128	0.15978	-0.22088	0.73193	-0.37862	0.77430
2	86	-0.79436	-0.18128	-1.02319	0.62456	-0.87649
3	12	-0.93105	-2.51375	-0.31450	-1.95620	1.33677
4	74	0.79778	1.00037	-0.02600	0.21386	-0.53748

adding the mean for each variable. In addition, PROC FASTCLUST could generate a list of observed individuals and their belongings to one of the four clusters.

16.4 Variable Cluster

When we want to group variables into different clusters, we use their correlation coefficients as the index of closeness. For example, in human genetics, people from different ethnicity and/or countries are studied. The individuals can be the genotypes and the cluster analysis will cluster the genotypes based on the "distance" between people's genetic profiles. The SAS procedure PROC VARCLUS is a tool to perform such analysis. Its principle is the same as the FASTCLUS. Once we standardized the variables, we can treat these variables as "individuals" and calculate their "distance" according to their similarity, which is measured by correlation coefficients. Once we have a cluster of more than one variable, we take average, either weighted or unweighted, of these variables in the cluster. These representative averages are the "centroids" of variables in their corresponding clusters. We can then perform the variable cluster again on these averages by treating them as new variables. Based on this principle, it is not difficult to understand the following processes.

- (1) To set initial seeds by assigning the number of clusters k ;
- (2) To calculate the representative averages;
- (3) To assess the closeness of the k representative averages and update the number of clusters, k_1 ;
- (4) To replace k_1 as k in (1) and repeat steps (2) and (3), until no more improvement exists.

In SAS PROC VARCLUS, selecting option CENTROID means to use simple (unweighted) average of all variables in a class as its representative average. If no option is selected, the principal factor is taken as the representative average. Also, there are options to limit the computation, such as MAXC that specifies the maximum number of clusters. If MAXC is not chosen, the maximum number of clusters is the number of variables. PROPORTION specifies the minimum percentage of the variance of cluster among the overall total variance, which also eliminates the small clusters. If it is not chosen, the minimum proportion is 1. MAXEIGEN specifies the allowance of the second eigenvalues. Beyond this level, the cluster should be broken into two clusters. This will assure the principal factor to represent the cluster. MAXEIGEN is 1 when PROPORTION and MAXC are not specified and 0 otherwise. When the method of centroid is chosen, you should not use the MAXEIGEN, because the centroid uses the average and the latter uses the principal factor.

PROC VARCLUS can be used to develop cluster levels like hierarchical cluster analysis by choosing the specification of HIERARCHY. The output can be obtained using OUTTREE and a specification of an SAS dataset. Then PROC TREE can be used to plot the tree diagram.

Example 16.3 Using data in Example 16.1 to perform a variable cluster analysis.

Solution Using PROC VARCLUS without selecting specific options, it will invoke the default principal factor method and the second eigenvalue cannot be larger than 1. After clustering, the 12 bacteria family becomes three clusters: X4, X5, X8 and X11 are in the first cluster, X1, X3, X7, X9, X10, and X12 are in the second cluster, X2 and X6 are in the third cluster.

16.5 Computerized Experiments

Experiment 16.1 Hierarchical cluster analysis Using the data in Example 16.1, we want to run a hierarchical cluster analysis based on complete linkage, single linkage, average linkage, flexible-beta, and Ward's minimum-variance methods. We want to compare the cluster results.

Program 16.1 Hierarchical cluster using flexible-beta method.

Line	Program
01	DATA A;
02	INPUT NAME \$ X1-X12;
03	CARDS;
04	HP1 0.12 25.42 0.00 7.72 0.00 0.00 0.00 29.06 25.92 0.00 11.76 0.00
05	XX4 3.85 6.76 0.19 38.95 10.10 0.00 12.24 2.47 18.95 0.00 6.40 0.10
06	;
07	PROC CLUSTER DATA=A OUTTREE=B METHOD=FLEXIBLE; VAR X1-
08	X12;VAR X1-X12;
09	ID NAME;
10	RUN;
11	PROC TREE DATA=B PAGES=1 SPACES=1;
12	ID NAME;
13	RUN;

Lines 01 to 06 in Program 16.1 read the data into the SAS dataset A. The variable NAME is the assigned indicator of bacteria types. Line 7 calls the program PROC CLUSTER. In the option, we selected OUTTREE=B to output cluster analysis results to SAS data file B. We selected METHOD=FLEXIBLE to suggest the use of flexible beta method. Other possible choices include COMPLETE for complete linkage, SINGLE for single linkage, AVERAGE for average linkage, and WARD for Ward's minimum variance, etc. The choice of STD standardizes the variables by subtracting the mean and then divided by the standard deviation. Line 08 lists the variables to be clustered, i.e., X1 to X12. Line 09 specifies using of bacteria type (NAME) as the identification (ID) in the output. Without this specification, the results would be named as OB1, OB2, . . . ,OB12. Line 11 of the program calls the PROC TREE to print the cluster analysis results. By selecting PAGES=1, the size of the tree diagram is limited to one page. Selection of SPACES=1 made the space between two individuals by one empty line. Usually, the tree is displayed in a vertical format with roots on the top and leaves at the bottom. If we add the option HORIZONTAL, the tree will be displayed in a horizontal format with root on the left and leaves on the right.

Experiment 16.2 Fast cluster analysis Using data in Example 2 of Appendix III to classify the normal ears into no more than four clusters.

Program 16.2 Fast cluster analysis.

Line	Program
01	DATA A;
02	INPUT EC 2.1 EK 2.1 EZ 2.1 EX 1.0 ECX 1.0;
03	EI=EK/EC*100; AI=EZ/EK*100;
04	CARDS;
05	663519531
06	653215524
07	;
08	PROC STANDARD DATA=A OUT=B MEAN=0 STD=1;
09	VAR EC EK EZ EI AI;
10	PROC FASTCLUS DATA=B MEAN=C1 OUT=C2 MAXC=4 DISTANCE;
11	VAR EC EK EZ EI AI;
12	RUN;
13	PROC PRINT DATA=C1;
14	PROC PRINT DATA=C2;
15	RUN;

Experiment 16.3 Variable cluster analysis Using the data in Experiment 16.1 to perform a variable cluster analysis. In this example, we use VARCLUS to cluster the 12 fatty acid variables. We did not select any options and therefore, the default principal factor approach was used with the default constraints that the second eigenvalue within a cluster should not be more than 1.

Program 16.3 Variable cluster analysis.

Line	Program
01	DATA A;
02	INPUT NAME \$ X1-X12;
03	CARDS;
04	HP1 0.12 25.42 0.00 7.72 0.00 0.00 0.00 29.06 25.92 0.00 11.76 0.00
05	XX4 3.85 6.76 0.19 38.95 10.10 0.00 12.24 2.47 18.95 0.00 6.40 0.10
06	;
07	PROC VARCLUS DATA=A;
08	RUN;

Chapter 17

Principal Component Analysis

When several parameters are measured from a subject, multiple random variables are involved, which reflect different aspects of the subject. Although there is no preference among these variables when they are collected, the amount of information provided by these variables are not necessary the same. If we do not know how to summarize these variables, the principal component analysis can be used.

17.1 The Basic Concepts of Principal Component Analysis

17.1.1 *Searching for the summary variable*

Example 17.1 The development indices of urban young adults

Table 17.1 listed the average values of several development indices of young adults of Han ethnicity obtained by 1985 Chinese Health Census, aged 19 to 22, from 28 provinces and cities in China with the numbers 1, 2, ..., 28 representing Beijing, Tianjing, Hebei, ..., etc., Hainan was part of Guangdong at that time. There were no data from Tibet in this table.

We consider the data in Table 17.1 as a representative sample of past and future of similar measurements in young adults. The six indices varied among provinces and their pairwise correlations were different. We would want to have a single or a few variables to summarize most of the information provided by the six development indices. Thus, we can use one or few variables instead of all six variables in application.

In addition to reliability and accuracy, a useful index for development should have variation ranges to reflect differences in the population. If an index is always the same for different individual subjects, it does not provide any information to separate them. Therefore, an index with a large between-subject variation is a good summary index.

Table 17.1 Average development indices of Han young adults (Age 19–22) in 1985 Chinese census.

Province #	Height (cm) X_1	Sitting height (cm) X_2	Weight (kg) X_3	Chest size (cm) X_4	Shoulder width (cm) X_5	Pelvic width (cm) X_6
1	173.28	93.62	60.10	86.72	38.97	27.51
2	172.09	92.83	60.38	87.39	38.62	27.82
3	171.46	92.78	59.74	85.59	38.83	27.46
4	170.08	92.25	58.04	85.92	38.33	27.29
5	170.61	92.36	59.67	87.46	38.38	27.14
6	171.69	92.85	59.44	87.45	38.19	27.10
7	171.46	92.93	58.70	87.06	38.58	27.36
8	171.60	93.28	59.75	88.03	38.68	27.22
9	171.60	92.26	60.50	87.63	38.79	26.63
10	171.16	92.62	58.72	87.11	38.19	27.18
11	170.04	92.17	56.95	88.08	38.24	27.65
12	170.27	91.94	56.00	84.52	37.16	26.81
13	170.61	92.50	57.34	85.61	38.52	27.36
14	171.39	92.44	58.92	85.37	38.83	26.47
15	171.83	92.79	56.85	85.35	38.58	27.03
16	171.36	92.53	58.39	87.09	38.23	27.04
17	171.24	92.61	57.69	83.98	39.04	27.07
18	170.49	92.03	57.56	87.18	38.54	27.57
19	169.43	91.67	57.22	83.87	38.41	26.60
20	168.57	91.40	55.96	83.02	38.74	26.97
21	170.43	92.38	57.87	84.87	38.78	27.37
22	169.88	91.89	56.87	86.34	38.37	27.19
23	167.94	90.91	55.97	86.77	38.17	27.16
24	168.82	91.30	56.07	85.87	37.61	26.67
25	168.02	91.26	55.28	85.63	39.66	28.07
26	167.87	90.96	55.79	84.92	38.20	26.53
27	168.15	91.50	54.56	84.81	38.44	27.38
28	168.99	91.52	55.11	86.23	38.30	27.14

17.1.2 Definition of the principal components

First, we should standardize all variables. Let \bar{X}_i and S_i be the sample mean and sample standard deviation for $i = 1, 2, \dots, 6$. Let

$$x_{ij} = \frac{X_{ij} - \bar{X}_i}{S_i} \quad (17.1)$$

16.6 Practice and Experiments

1. Both the cluster analysis and discriminant analysis are the classification techniques for multivariate data analysis. How are they different? What kind of problems do they solve respectively? Give examples to illustrate your points.
2. If we have already known the cluster of Example 16.1 results that HP bacteria form one cluster; CJ and CC are in the second cluster; and the rest four enterobacteria forms the third cluster. Using X1–X12 to perform a discriminant analysis. Are the discriminant results reasonable?
3. Using the data in Example 16.1 and perform a fast cluster analysis with $MAXC=2, 3$, and 4. Compare the results with Figs. 16.2 and 16.3 to see their differences from the clusters by average linkage method in the cluster sizes of 2, 3 and 4.
4. Using the variables of EC, EX, EZ, EI and AI in Example 16.2 to perform a hierarchical cluster analysis via centriod method. Through this exercise, how do you think about the use of hierarchical cluster analysis for large sample data?
5. In Chap. 14, we discussed discrimination analysis of retinopathy severity based on the data from an electro-retinography. Use the training samples with ten variables in the data set to perform a fast cluster analysis to form three severity groups. Compare the severity groups with the three groups given in the training data and see if they agree with each other.
6. An orthodontic clinic wanted to study the erupted wisdom tooth. 50 patients with early erupted wisdom tooth were used as study individuals. Every patient had a head X-ray film and from which 25 quantitative measurements were derived. Data are given in Appendix III.
 - (1) Try to use these 25 quantitative measures to define hierarchical clusters of individuals using different distance metrics. Cluster the data based on the tree diagram to arrange the 50 patients into three clusters. Compare the cluster results against the classification results by experts based on morphology of bone and tooth. Which distance metric is the best method to generate clusters that is most similar to those suggested by orthodontic experts?

- (2) Use 25 measures, various distance metrics, and initial cluster seeds based on experts' classifications to perform fast cluster analyses of 50 patients. The maximum number of clusters is fixed at 3. Compare your cluster results with the orthodontist classifications.

(1st edn. Ying Lu; 2nd edn. Yuanto Hao, Weiyang Su, Jiqian Fang)

with $j = 1, 2, \dots, 28$, representing the averages from 28 provinces and cities. Obviously, the sample mean and sample standard deviation of x_{ij} were 0 and 1, respectively. We use $x_i, i = 1, 2, \dots, 6$, to denote the standardized variables. We can search for summary indices one by one by the following steps.

(1) Denote the first summary index as C_1 . C_1 is selected from linear combination of x_1, x_2, \dots, x_6 . That is

$$C_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{16}x_6. \quad (17.2)$$

We do not care about the scales of $a_{11}, a_{12}, \dots, a_{16}$ as long as they change proportionally. The between-subject variation is proportional to the scales. Thus, we can simply put the constrain

$$a_{11}^2 + a_{12}^2 + a_{13}^2 + a_{14}^2 + a_{15}^2 + a_{16}^2 = 1. \quad (17.3)$$

To make C_1 the best, we want it to have the maximum variation among samples than any other linear combinations. That is $Var(C_1)$ is maximized among all linear combinations with constrains of (17.3).

(2) We can also search for the second best index. It should be another linear combination of x_1, x_2, \dots, x_6 .

$$C_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{26}x_6. \quad (17.4)$$

In order to effectively represent information from original variables, there is no need to include any information represented by C_1 in the second index C_2 . Thus, besides the constraints of (17.3), the vector of coefficients for C_2 , denoted by $(a_{21}, a_{22}, \dots, a_{26})$, should be orthogonal to the vector of coefficients for C_1 . There are many such choices and we want to use the best linear combination such that $Var(C_2)$ is maximized among all possible choices.

(3) We can continue to search for the next summary index. As there are only six original variables, the maximum independent indices is not more than six.

In general, suppose that we have p random variables X_1, X_2, \dots, X_p , and have obtained their sample means $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p$ and sample standard deviations S_1, S_2, \dots, S_p . After performing standardization as in (17.1), we

have the following definitions:

- (1) For all linear combinations

$$C_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$$

with the constraint that

$$a_{11}^2 + a_{12}^2 + \cdots + a_{1p}^2 = 1$$

the first principal component is the linear combination that has maximum variance $\text{Var}(C_1)$.

- (2) For all linear combinations

$$C_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p$$

with the constraints that

$$a_{21}^2 + a_{22}^2 + \cdots + a_{2p}^2 = 1 \quad \text{and} \quad \sum_{i=1}^p a_{1i}a_{2i} = 0$$

(i.e., the coefficients for C_1 and C_2 are orthogonal to each other), the second principal component is the linear combination that has maximum variance $\text{Var}(C_2)$.

- (3) We can similarly define the third, fourth, ..., up to p th principal components.

17.1.3 Properties of principal components

Principal components C_1, C_2, \dots, C_p have the following properties.

- (1) Principal components are uncorrelated. That is, for any pair i and j , the correlation coefficient between C_i and C_j is zero.

$$\text{Corr}(C_i, C_j) = 0. \quad (17.5)$$

- (2) Coefficients $(a_{i1}, a_{i2}, \dots, a_{ip})$ are unit vector, i.e.,

$$a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 = 1. \quad (17.6)$$

- (3) The variances of principal components decrease, i.e.,

$$\text{Var}(C_1) \geq \text{Var}(C_2) \geq \cdots \geq \text{Var}(C_p). \quad (17.7)$$

- (4) The total variance of principal components is the same as the total variance of the original variables, i.e.,

$$\begin{aligned}
 & \text{Var}(C_1) + \text{Var}(C_2) + \cdots + \text{Var}(C_p) \\
 &= \text{Var}(x_1) + \text{Var}(x_2) + \cdots + \text{Var}(x_p) \\
 &= p.
 \end{aligned} \tag{17.8}$$

This property reveals that principal components are formed by reorganizing the original variables. They do not gain or lose any information from the original variables.

17.2 Computation and Interpretation of Principal Components

17.2.1 Computational procedures

Although from their definitions, the principal components seem to be calculated stepwise, the actual computation procedure is not stepwise.

Let R be the correlation matrix of X_1, X_2, \dots, X_p .

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{pmatrix}. \tag{17.9}$$

We can prove that the combination coefficients for the i th principal component $a_{i1}, a_{i2}, \dots, a_{ip}$ are the solution for the following linear equations based on the definition of principal components.

$$\begin{aligned}
 (1 - \lambda_i)a_{i1} + r_{12}a_{i2} + \cdots + r_{1p}a_{ip} &= 0, \\
 r_{12}a_{i1} + (1 - \lambda_i)a_{i2} + \cdots + r_{2p}a_{ip} &= 0, \\
 &\vdots \\
 r_{1p}a_{i1} + r_{2p}a_{i2} + \cdots + (1 - \lambda_i)a_{ip} &= 0.
 \end{aligned} \tag{17.10}$$

Here, λ_i is the i th eigenvalue of the correlation matrix R ,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0 \tag{17.11}$$

Table 17.2 Correlation matrix of data in Table 17.1.

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1.0000	0.9557	0.8539	0.4140	0.1815	0.1004
X_2		1.0000	0.8073	0.4041	0.2471	0.2362
X_3			1.0000	0.5326	0.2416	0.0581
X_4				1.0000	-0.0541	0.3302
X_5					1.0000	0.4358
X_6						1.0000

and $(a_{i1}, a_{i2}, \dots, a_{ip})$ is the corresponding eigenvector. Furthermore, the eigenvalue λ_i is the variance of the i th principal component. All the correlation coefficients between principal components and initial variables for Example 17.1 are given in Table 17.5.

$$\text{Var}(Ci) = \lambda_i. \quad (17.12)$$

Thus, we can use the following procedures to calculate the principal components:

- (1) Calculate the correlation matrix R of X_1, X_2, \dots, X_p .
- (2) Calculate the eigenvalues of R .
- (3) Calculate the corresponding eigenvectors for the eigenvalues.

Example 17.2 Calculate all principal components according to the data in Table 17.1.

Solution First we obtain the correlation matrix as in Table 17.2. Based on the matrix in Table 17.2, we can obtain the eigenvalues in Table 17.3.

The second column in Table 17.3 presents the i th eigenvalue, which is the variance of the i th principal component, $i = 1, 2, \dots, 6$. The third column shows the differences between the variances of two consecutive principal components, i.e., the magnitudes of decreases. The fourth column reflects the contribution of i th principal component, and the fifth column, the contribution up to the i th principal component, $i = 1, 2, \dots, 6$. The coefficients for the principal components, i.e., the eigenvectors, are given in Table 17.4.

Table 17.3 Eigenvalues of correlation matrix for data in Example 17.1.

Principal components	Eigenvalues	Differences between two eigenvalues	Percentage variations	Cumulative percentage variations
1st	3.17310	1.85631	0.52885	0.52885
2nd	1.31678	0.38000	0.21946	0.74831
3rd	0.93679	0.51650	0.15613	0.90445
4th	0.42028	0.29894	0.07005	0.97449
5th	0.12135	0.08965	0.02023	0.99472
6th	0.03170		0.00528	1.00000

Table 17.4 Eigenvectors for the eigenvalues in Example 17.1.

Variables	1 st	2 nd	3 rd	4 th	5 th	6 th
X_1	<u>0.522252</u>	-0.195699	-0.189953	-0.253741	0.226568	0.732908
X_2	<u>0.525559</u>	-0.080164	-0.167681	-0.388390	0.304015	-0.667812
X_3	<u>0.511208</u>	-0.181857	-0.103986	0.334729	-0.758103	-0.089540
X_4	0.345993	-0.046978	<u>0.741653</u>	0.456060	0.346103	-0.015969
X_5	0.188783	<u>0.656595</u>	-0.470338	0.498021	0.252640	0.013219
X_6	0.185358	<u>0.699199</u>	0.392072	-0.465521	-0.312900	0.091793

17.2.2 The number of principal components

We have previously pointed out that p random variables produce p principal components. As the total variance remains the same for the original variables and the principal components, the first few principal components, such as C_1 and C_2 , account for more variation than the last few principal components, such as C_{p-1} and C_p . In our example in Table 17.3, $\text{Var}(C_1) = 3.17310$, $\text{Var}(C_2) = 1.31678$, but $\text{Var}(C_6) = 0.03170$, $\text{Var}(C_5) = 0.12135$. Therefore, only the first few components play the role of "principal" ones. We always keep the first few principal components and drop the rest in practice.

How many principal components to keep and how much should be sufficient depend on the cumulative variance explained by these components. In applications, we can roughly decide the percentage of variance we want to retain and thus to decide how many principal components to keep. If retaining a principal component adds only slightly to the cumulative variance,

this principal component should be dropped. In Table 17.3, the first three principal components account for 90.45% total variation, which is the majority of information we are interested in so that keeping the first three principal components is appropriate.

Another general rule is that a principal component is worth considering if the corresponding eigenvalue is above one. Using this rule to the above example, the third principal component should be dropped because its eigenvalue is less than 1. If that is the case, the first two principal components explained 74.8% of the total variance. Therefore, how many principal components to keep depends on the satisfaction of the researchers with the information explained by principal components.

17.2.3 Correlation between principal components and variables

Let C_i be the i th principal component,

$$C_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ip}x_p. \quad (17.13)$$

The relationship between C_i and x_j can be described by their correlation coefficient:

$$\text{Corr}(C_i, x_j) = a_{ij} \sqrt{\text{Var}(C_i)} = a_{ij} \sqrt{\lambda_i}. \quad (17.14)$$

In Table 17.4, we underlined the j th element of the i th eigenvector that the absolute value of the correlation coefficient $\text{Corr}(C_i, x_j)$ is greater than 0.5. For example

$$\begin{aligned} \text{Corr}(C_1, x_1) &= 0.522252\sqrt{3.17310} = 0.930298, \\ \text{Corr}(C_2, x_1) &= -0.195699\sqrt{1.31678} = -0.224567. \end{aligned}$$

Thus, the first element of the first eigenvector (0.522252) was underlined, but the first element of the second eigenvector (-0.195699) was not.

There are no eigenvectors underlined for the fourth to sixth principal components because their correlations with original variables are all small.

For the first three principal components, the underlined elements of the eigenvectors suggest stronger correlation between the principal components and these variables. It helps to interpret the principal components. For example, the first principal component mainly reflects the body size because it correlates highly with the variables height, sitting height, weight, and chest

Table 17.5 Correlation between principal components and variables.

	C_1	C_2	C_3	C_4	C_5	C_6
X_1	0.9303	-0.2246	-0.1839	-0.1645	0.0789	0.1305
X_2	0.9362	-0.0920	-0.1623	-0.2518	0.1059	-0.1189
X_3	0.9106	-0.2087	-0.1007	0.2170	-0.2641	-0.0159
X_4	0.6163	-0.0539	0.7178	0.2957	0.1206	-0.0028
X_5	0.3363	0.7535	-0.4552	0.3229	0.0880	0.0024
X_6	0.3302	0.8023	0.3795	-0.3018	-0.1090	0.0163

size. The second principal component correlates more with the shoulder width and pelvic width and hence reflects the body width. However, not all components can be easily interpreted, such as the fourth to the sixth principal components.

In summary, the principal component analysis is useful to summarize data. However, not all components have clear interpretations.

17.3 Principal Component Analysis in Regression

Principal component analysis is often not an end in itself, but rather a means to achieve the goals. They can be used as the intermediate step in multivariate regression, cluster analysis and discriminant analysis, etc. The goal of this subsection is to introduce the principal component analysis in regression.

As we pointed out in the previous chapter of multiple regression analysis, the regression coefficients can be highly unstable when the independent variables are strongly correlated. Sometimes, we have uninterpretable counter intuitive results. In such cases, we can use the principal component analysis to reorganize the original variables into several principal components. The strongly correlated variables will most likely be represented by the same principal component and different principal components are uncorrelated. As long as we keep enough principal components, we will retain most information of the original variables. Using the principal components as independent variables in regression analysis will avoid the problem of collinearity. When we drop the last few principal components in the regression model, we can get reasonable results. Of course, if we use all p principal components, the regression will be the same as the use of all original variables.

Example 17.3 Dr. Chen (Chinese Health Statistics, Vol. 8, Issue 1, 1991) reported an example of principal component analysis in regression. In the study, 22 fetuses were observed for their age since fertilization (Y weeks) and morphology, such as their height (X_1 , cm), head circumference (X_2 , cm), and weight (X_3 , g). The goal was to derive a regression equation of Y on variables X_1 , X_2 and X_3 . The data are given in Table 17.6.

Solution If we directly perform a regression analysis using independent variables X_1 , X_2 and X_3 , we had the following regression equation:

$$\hat{Y} = 11.0117 + 1.6927X_1 - 2.1588X_2 + 0.0075X_3. \quad (17.15)$$

Table 17.6 The age since fertilization and the morphology measures of 22 fetuses.

No.	Height (cm) X_1	Head circumference (cm) X_2	Weight (g) X_3	Weeks since fertilization Y
1	13.0	9.2	50.0	13.0
2	18.7	13.2	102.0	14.0
3	21.0	14.8	150.0	15.0
4	19.0	13.3	110.0	16.0
5	22.8	16.0	200.0	17.0
6	26.0	18.2	330.0	18.0
7	28.0	19.7	450.0	19.0
8	31.4	22.5	450.0	20.0
9	30.3	21.4	550.0	21.0
10	29.2	20.5	640.0	22.0
11	36.2	25.2	800.0	23.0
12	37.0	26.1	1090.0	24.0
13	37.9	27.2	1140.0	25.0
14	41.6	30.0	1500.0	26.0
15	38.2	27.1	1180.0	27.0
16	39.4	27.4	1320.0	28.0
17	39.2	27.6	1400.0	29.0
18	42.0	29.4	1600.0	30.0
19	43.0	30.0	1600.0	31.0
20	41.1	27.2	1400.0	33.0
21	43.0	31.0	2050.0	35.0
22	49.0	34.8	2500.0	36.0

Table 17.7 The correlation matrix of Example 17.3.

	X_1	X_2	X_3	Y
X_1	1.0000			
X_2	0.9975	1.0000		
X_3	0.9441	0.9470	1.0000	
Y	0.9525	0.9430	0.9701	1.0000

Table 17.8 Principal component analysis of Example 17.3.

	C_1	C_2	C_3
X_1	0.58057	-0.41852	0.69841
X_2	0.58107	-0.38789	-0.71547
X_3	0.57034	0.82121	0.01799
Var	2.9261	0.0714	0.0025
% variation	97.54	2.38	0.08
Cumulative % variation	97.54	99.92	100

Here, the regression coefficient for head circumference was negative, which was counter intuitive as we expected the positive relationship between fertilization age and head circumference. It was easy to see the reason of such negative coefficients because of high correlation between the three independent variables as shown in Table 17.7. For example, the correlation coefficient between X_1 and X_2 was 0.9975. The other two correlation coefficients were also high. There was a strong colinearity of the three variables. Thus, the regression coefficients of X_1 and X_3 in (17.15) reflected not only their contribution to Y but also the contribution of X_2 to Y . Thus, a negative coefficient of X_2 was necessary to compensate the over contribution of X_1 and X_3 .

A principal component analysis was performed, and the results were summarized in Table 17.8.

As $Var(C_3) \approx 0$, we only kept the first two principal components C_1 and C_2 . The regression equation of Y on C_1 and C_2 was

$$\hat{Y} = 23.72727 + 3.882227C_1 + 3.099072C_2$$

with

$$\begin{aligned}
 C_1 &= 0.58057 \frac{X_1 - 33.04549}{9.710168} + 0.58107 \frac{X_2 - 23.26364}{6.857498} \\
 &\quad + 0.57034 \frac{X_3 - 936.9091}{690.3048}, \\
 C_2 &= -0.41852 \frac{X_1 - 33.04549}{9.710168} + 0.38789 \frac{X_2 - 23.26364}{6.857498} \\
 &\quad + 0.82121 \frac{X_3 - 936.9091}{690.3048}.
 \end{aligned}$$

Replace C_1 and C_2 by the above expressions, the new regression equation was

$$\hat{Y} = 10.43671 + 0.09854X_1 + 0.15366X_2 + 0.00689X_3$$

which had a positive regression coefficient for X_2 . The numerical confusion caused by the direct regression using X_1 , X_2 and X_3 was resolved.

Similar idea can be used for discriminant analysis. When the independent variables are strongly correlated, using them directly in discriminant analysis will face the same colinearity problem. The problem can be avoided by performing the principal component analysis first and then use the principal components to perform discriminant analysis. This approach is also applicable to in Chaps. 28 and 29 for logistic and Cox regression analysis.

17.4 Computerized Experiments

Experiment 17.1 Sampling experiment of principal component analysis

Let us generate a value from standard normal distribution $N(10, 1)$, denoted as X_1 , then generate a value from $N(2X_1, 1)$, denoted by X_2 ; further generate a value from $N(30, 4)$, denoted by X_3 , then generate a value from $N(3X_3, 4)$, denoted by X_4 ; gathering these four values as a vector. Repeating this experiment 100 times creates a “sample” of (X_1, X_2, X_3, X_4) with 100 “individuals”. Perform the following analyses:

- (1) Calculate the variance of four variables and the total variance;
- (2) Draw histograms of the four variables;

- (3) Calculate the correlation matrix of four variables and summarize their characteristics;
- (4) Calculate four principal components C_1 , C_2 , C_3 and C_4 , their contributions to total variance, and the variables they represent;
- (5) Calculate C_1 , C_2 , C_3 and C_4 for each individual;
- (6) Calculate the variances of C_1 , C_2 , C_3 and C_4 and their total variance. Compare the total variance with (1);
- (7) Draw histograms for C_1 , C_2 , C_3 and C_4 and compare them with (2);
- (8) Calculate correlation matrix of C_1 , C_2 , C_3 and C_4 and compare it with (3).

Program 17.1 Sampling experiment of principal component analysis.

Line	Program	Line	Program
01	DATA A;	17	VBAR X2/MIDPOINTS=15 TO 25 BY 0.5;
02	DO I=1 TO 100;	18	PROC GCHART;
03	X1=RANNOR(0)+10;	19	VBAR X3 / MIDPOINTS=24 TO 36 BY 1;
04	X2=RANNOR(0)+X1*2;	20	PROC GCHART;
05	X3=RANNOR(0)*2+30;	21	VBAR X4 / MIDPOINTS=84 TO 96 BY 1;
06	X4=RANNOR(0)*2+X3*3;	22	PROC PRINCOMP OUT=B PREFIX=C;
07	OUTPUT;	23	VAR X1 X2 X3 X4;
08	END;	24	PROC MEANS;
09	PROC MEANS;	25	VAR C1 C2 C3 C4;
10	VAR X1 X2 X3 X4;	26	PROC CORR;
11	PROC CORR;	27	VAR C1 C2 C3 C4;
12	VAR X1 X2 X3 X4;	28	PROC GCHART;
13	GOPTIONS DEVICE=WIN;	29	VBAR C1 / LEVELS=12;
14	PROC GCHART;	30	PROC GCHART;
15	VBAR X1 /MIDPOINTS=7	31	VBAR C2 /LEVELS=10;
	TO 13 BY 0.5;	32	RUN;
16	PROC GCHART;		

Lines 02 to 08 of Program 17.1 generate 100 “individuals” and assign them to dataset A. Lines 09 to 10 calculate the mean and standard deviations of the four variables. Lines 11 to 12 calculate correlation matrix. Lines 13 to 21 draw four histogram charts. Readers may adjust the midpoints and ranges to display a better figure. Lines 22 and 23 perform a principal component analysis and store the output into SAS dataset B. The levels are

specified as 12 and 10 respectively. Readers can adjust the levels to see the changes.

Experiment 17.2 Principal component analysis in regression of strongly correlated variables Let X_1 as a random value from $N(10, 1)$ and X_2 as random value from $N(X_1, 1)$; further generate X_3 as a random value from $N(X_1 - X_2, 1)$ and X_4 as a random value from $N(X_1 + X_2 - X_3, 1)$. Y is a random value from $N(X_1 + 2X_2 + 3X_3 + 4X_4, 1)$; gathering these four values as a vector. Repeating the “sampling” procedure 100 times generates a study “sample” of (X_1, X_2, X_3, X_4) .

- (1) Perform a regression of Y on X_1, X_2, X_3, X_4 ;
- (2) Perform a principal component analysis for X_1, X_2, X_3, X_4 , and keep some principal components to maintain 95% of total variations;
- (3) Perform a regression of Y on the retained principal components;
- (4) Replace the principal components by the original variables X_1, X_2, X_3, X_4 using the relationship between them;
- (5) Compare the results (1) and (4) with the expected relationship that $\mu_{Y|X} = X_1 + 2X_2 + 3X_3 + 4X_4$.

Program 17.2 Principal components analysis in multiple regression.

Line	Program	Line	Program
01	DATA A;	09	END;
02	DO I=1 TO 100;	10	PROC REG;
03	X1=RANNOR(0)+10;	11	MODEL Y=X1 X2 X3 X4;
04	X2=RANNOR(0)+X1;	12	PROC PRINCOMP OUT=B
05	X3=RANNOR(0)+X1-X2;		PREFIX=C;
06	X4=RANNOR(0)+X1+X2-X3;	13	VAR X1 X2 X3 X4;
07	Y=RANNOR(0)+X1+X2*2+	14	PROC REG;
	X3*3+X4*4;	15	MODEL Y=C1;
08	OUTPUT;	16	RUN;

The first half of the program is similar to Program 17.1, which generates Y and X_1, X_2, X_3, X_4 . In program 17.2, lines 10 and 11 perform a regression analysis. Lines 12 and 13 perform a principal component analysis. Lines

Table 17.9 Data of 25 controls and 25 CHD patients.

ID	Patients $G = 1$					ID	Controls $G = 2$				
	X_1	X_2	X_3	X_4	X_5		X_1	X_2	X_3	X_4	X_5
1	61	170	198	88	93	26	63	100	154	44	83
2	66	130	233	200	100	27	55	130	195	124	100
3	64	190	205	50	102	28	64	104	216	100	110
4	73	140	186	133	106	29	59	150	229	175	85
5	59	140	294	250	110	30	40	120	128	67	100
6	66	140	225	144	92	31	59	150	229	175	85
7	55	144	181	44	96	32	56	100	134	100	85
8	47	120	167	142	87	33	53	138	206	40	86
9	83	170	158	133	85	34	57	100	181	50	97
10	81	124	188	100	91	35	45	110	186	67	79
11	73	180	223	150	90	36	60	120	154	100	95
12	76	170	198	163	99	37	60	150	167	89	88
13	66	178	223	83	98	38	70	132	191	344	118
14	67	166	109	56	96	39	59	120	187	140	87
15	75	166	218	89	96	40	72	120	186	150	88
16	70	100	259	83	104	41	58	150	155	89	72
17	75	176	233	167	90	42	58	130	124	78	75
18	71	120	179	100	168	43	41	120	217	344	111
19	75	230	174	67	157	44	47	90	184	74	90
20	66	176	191	80	88	45	62	146	134	56	85
21	61	156	178	200	97	46	69	150	211	100	93
22	72	170	160	100	88	47	45	96	131	78	108
23	63	140	198	78	100	48	61	130	163	67	94
24	73	150	212	122	190	49	53	100	183	50	94
25	58	150	132	150	95	50	80	170	165	67	104

14 to 16 perform a regression analysis on the first principal component. Readers should try to add second and third principal components.

17.5 Practice and Experiments

1. Pan (Chinese Health Statistics, 9(4), 1993) provided a data set of 25 normal controls and 25 patients with cardiovascular heart disease (CHD), which is now given in Table 17.9.

In Table 17.9, X_1 is age in years; X_2 is systolic blood pressure in unit of mmHg, X_3 is cholesterol level in unit of mg/100ml; X_4 is triglyceride level in unit of mg/100ml.

Table 17.10 Observed data of 15 male infants.

ID	X_1 (cm)	X_2 (kg)	Y (cm ²)
1	54.0	3.00	2446.2
2	50.5	2.25	1928.4
3	51.0	2.50	2094.5
4	56.5	3.50	2506.7
5	52.0	3.00	2121.0
6	76.0	9.50	3845.9
7	80.0	9.00	4380.8
8	74.0	9.50	4314.2
9	80.0	9.00	4078.4
10	76.0	8.00	4134.5
11	96.0	13.50	5830.2
12	97.0	14.00	6013.6
13	99.0	16.00	6410.6
14	92.0	11.00	5283.3
15	94.0	15.00	6101.6

- (1) Perform a binary discriminant analysis to generate a discriminant function;
- (2) Perform a principal components analysis and use the principal components to generate a discriminant function; compare the results with (1).

2. Zhen and Wang (Chinese Health Statistics; 11(3), 1994) reported a dataset of height (X_1 , cm), weight (X_2 , kg), and body surface area (Y , cm²) from 30 infants. We use a subset of it with 15 male infant data in Table 17.10.

- (1) Make a scatter plot of X_1 and X_2 ;
- (2) Perform a principal component analysis for X_1 and X_2 ;
- (3) Draw the vector of principal components in the scatter plot (1), and discuss the interpretation of the two principal components;
- (4) Perform a regression analysis of Y on the first principal component;
- (5) Perform a regression analysis of Y on the second principal component;
- (6) Perform a regression analysis of Y on both principal components;
- (7) Compare the regression results of (4), (5) and (6).

3. Using the 1995 functional data of Han ethnicity young adults aged 19–22 from 28 Chinese provinces/cities displayed in Table 17.11 to perform a principal component analysis.

Table 17.11 Functional data of Han ethnicity young adults aged 19–22.

ID of province/ cities	Pulse Y_1 (/min)	Systolic BP Y_2 (mmHg)	Diastolic BP (Sounds muffled) Y_3 (mmHg)	Diastolic BP (Sounds disappear) Y_4 (mmHg)	Lung capacity Y_5 (ml)
1	75.3	117.4	61.8	61.8	4508
2	76.7	120.1	66.2	66.2	4469
3	75.8	121.8	65.4	65.4	4398
4	76.1	115.1	61.3	61.3	4068
5	72.9	119.4	67.1	67.1	4339
6	72.7	116.2	59.3	59.3	4393
7	76.5	117.9	68.3	68.3	4389
8	75.2	115.1	63.2	63.2	4306
9	74.7	117.4	68.3	68.3	4395
10	73.2	113.2	51.0	51.0	4462
11	77.8	116.9	65.6	65.6	4181
12	76.4	113.6	65.6	65.6	4232
13	76.4	116.7	61.2	61.2	4305
14	74.9	113.1	61.2	61.2	4276
15	78.7	112.4	61.4	61.4	4067
16	73.9	118.4	62.3	62.3	4421
17	75.7	116.3	51.8	51.8	4284
18	72.5	114.8	55.1	55.1	4289
19	76.7	117.5	51.6	51.6	4097
20	77.0	117.9	52.4	52.4	4063
21	76.0	116.8	58.0	58.0	4334
22	74.2	115.4	60.4	60.4	4301
23	76.2	110.9	56.8	56.8	4141
24	77.2	113.8	57.5	57.5	3905
25	74.5	117.2	63.8	63.8	3943
26	74.3	112.3	50.2	50.2	4195
27	77.5	117.4	63.6	63.6	4039
28	77.7	113.3	52.8	52.8	4238

(1st edn. Ying Lu, Jie Yan; 2nd edn. Yuanto Hao, Yan Chen, Jiqian Fang)



Chapter 18

Factor Analysis

Chapter 17 introduces the principal component analysis that organizes the original variables into several principal components via linear combination of the original variables. We can use fewer number of principal components to retain the most information of the original variables. In this chapter, we will define factors to decompose original variables into several more clearly defined common components and, thus, to better understand the relationship among original variables.

18.1 Factor Model

Let X_1, X_2, \dots, X_p be p random variables with sample means $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p$ and sample standard deviations S_1, S_2, \dots, S_p . For $i = 1, 2, \dots, p$ let x_i be the standardized variable, i.e.,

$$x_i = \frac{X_i - \bar{X}_i}{S_i}. \quad (18.1)$$

Suppose there are m ($m \leq p$) different variables F_1, F_2, \dots, F_m with zero means and unit variances, i.e., for $i = 1, 2, \dots, m$

$$\bar{F}_i = 0, \quad \text{Var}(F_i) = 1, \quad (18.2)$$

F_1, F_2, \dots, F_m are linear combinations of X_1, X_2, \dots, X_p and should contain most information of them. These F_1, F_2, \dots, F_m are called common factors of X_1, X_2, \dots, X_p .

We want to express x_i in terms of F_1, F_2, \dots, F_m . Thus, x_i can be expressed as the linear combination of these common factors plus information of individual variables that are not explained by these common factors. Let individual factor be e_i . By definition, e_i is uncorrelated with

F_1, F_2, \dots, F_m . We have the following linear equations:

$$\begin{aligned} x_1 &= L_{11}F_1 + L_{12}F_2 + \dots + L_{1m}F_m + e_1, \\ x_2 &= L_{21}F_1 + L_{22}F_2 + \dots + L_{2m}F_m + e_2, \\ &\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ x_p &= L_{p1}F_1 + L_{p2}F_2 + \dots + L_{pm}F_m + e_p. \end{aligned} \quad (18.3)$$

L_{ij} is the factor loading of the variable x_i on factor F_j . Therefore, the variance of x_i can be expressed as follows:

$$Var(x_i) = Var(L_{i1}F_1 + L_{i2}F_2 + \dots + L_{im}F_m) + Var(e_i). \quad (18.4)$$

As the left of above equation is 1, it can be expressed as the sum of common variance (for commonality) and specific variance (for variable uniqueness). Let

$$h_i^2 = Var(L_{i1}F_1 + L_{i2}F_2 + \dots + L_{im}F_m), \quad (18.5)$$

$$u_i^2 = Var(e_i). \quad (18.6)$$

Thus

$$h_i^2 + u_i^2 = 1. \quad (18.7)$$

Based on this decomposition, we have to solve the following three tasks in a factor analysis

- (1) Find the common factors $F_1, F_2, \dots, F_m (m \leq p)$;
- (2) Calculate the factor loadings L_{ij} ;
- (3) Interpret relationships among the original variables.

18.2 Derivation of Factors

There are two commonly used methods for deriving factors: the principal component method and iterative principal component method.

18.2.1 Principal component method

Using principal component analysis, we can get p principal components

$$\begin{aligned} C_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p, \\ C_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p, \\ &\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ C_p &= a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p \end{aligned} \quad (18.8)$$

with $a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 = 1$ and for any pair $i \neq j$, $\sum_{k=1}^p a_{ik}a_{jk} = 0$.

When we view (18.8) as simultaneous linear equations, we can obtain the inverse solution that

$$\begin{aligned} x_1 &= a_{11}C_1 + a_{21}C_2 + \cdots + a_{p1}C_p, \\ x_2 &= a_{12}C_1 + a_{22}C_2 + \cdots + a_{p2}C_p, \\ &\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ x_p &= a_{1p}C_1 + a_{2p}C_2 + \cdots + a_{pp}C_p. \end{aligned} \quad (18.9)$$

Here, the coefficient matrix is the transpose of the coefficient matrix of (18.8). After standardizing the principal components C_1, C_2, \dots, C_p , we have the initial factors

$$F_j = \frac{C_j}{\sqrt{\text{Var}(C_j)}} \quad j = 1, 2, \dots, p. \quad (18.10a)$$

Obviously, $\bar{F}_j = 0$, $\text{Var}(F_j) = 1$.

Substituting

$$C_j = F_j \sqrt{\text{Var}(C_j)} \quad (18.10b)$$

into Eq. (18.9), we have

$$\begin{aligned} x_1 &= (a_{11}\sqrt{\text{Var}(C_1)})F_1 + (a_{21}\sqrt{\text{Var}(C_2)})F_2 + \cdots \\ &\quad + (a_{m1}\sqrt{\text{Var}(C_m)})F_m + \cdots, \end{aligned}$$

$$\begin{aligned}
x_2 &= (a_{12}\sqrt{\text{Var}(C_1)})F_1 + (a_{22}\sqrt{\text{Var}(C_2)})F_2 + \cdots \\
&\quad + (a_{m2}\sqrt{\text{Var}(C_m)})F_m + \cdots, \\
&\quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\
x_p &= (a_{1p}\sqrt{\text{Var}(C_1)})F_1 + (a_{2p}\sqrt{\text{Var}(C_2)})F_2 + \cdots \\
&\quad + (a_{mp}\sqrt{\text{Var}(C_m)})F_m + \cdots.
\end{aligned} \tag{18.11}$$

We denote

$$L_{ij} = a_{ji}\sqrt{\text{Var}(C_j)} \quad i = 1, 2, \dots, p; \quad j = 1, 2, \dots, m \tag{18.12}$$

and make all terms after m as e_1, e_2, \dots, e_p , then we can obtain the factor model in (18.3).

As expressed in (17.14) in Chap. 17, L_{ij} is the correlation coefficient between the i th variable and the j th principal component:

$$L_{ij} = \text{Corr}(X_i, C_j) = \text{Corr}(X_i, F_j).$$

Using (18.5) and (18.6), we have

$$\begin{aligned}
h_i^2 &= \text{Var}(L_{i1}F_1 + L_{i2}F_2 + \cdots + L_{im}F_m) \\
&= L_{i1}^2 \text{Var}(F_1) + L_{i2}^2 \text{Var}(F_2) + \cdots + L_{im}^2 \text{Var}(F_m) \\
&= L_{i1}^2 + L_{i2}^2 + \cdots + L_{im}^2 \\
&= a_{1i}^2 \text{Var}(C_1) + a_{2i}^2 \text{Var}(C_2) + \cdots + a_{mi}^2 \text{Var}(C_m), \\
e_i^2 &= 1 - h_i^2.
\end{aligned}$$

Example 18.1 (Continued from 17.1) Using the development indices of 1985 Chinese Han young adults aged 19–22 from 28 provinces/cities to perform a factor analysis.

Solution From Example 17.1, we have information of principal components and eigenvalues, which we display here again in Table 18.1.

Using formula (18.10a) and Table 18.1, we can derive coefficients for common factors, displayed in Table 18.2 (read in columns), and the common factors in Table 18.3 (read in rows).

Table 18.1 Coefficients (in columns) of four principal components and their variances.

	C_1	C_2	C_3	C_4
x_1	<u>0.522252</u>	-0.195699	-0.189953	-0.253741
x_2	<u>0.525559</u>	-0.080164	-0.167681	-0.388390
x_3	<u>0.511208</u>	-0.181857	-0.103986	0.334729
x_4	0.345993	-0.046978	<u>0.741653</u>	0.456060
x_5	0.188783	<u>0.656595</u>	-0.470338	0.498021
x_6	0.185358	<u>0.699199</u>	0.392072	-0.465521
Var	3.17310	1.31678	0.93679	0.42028

Table 18.2 Coefficients of common factors.

	F_1	F_2	F_3	F_4
x_1	0.29318	-0.17054	-0.19626	-0.3914
x_2	0.29504	-0.06986	-0.17325	-0.59910
x_3	0.28698	-0.15848	-0.10744	0.51633
x_4	0.19423	-0.04094	0.76627	0.70348
x_5	0.10598	0.57219	-0.48595	0.76821
x_6	0.10406	0.60932	0.40508	-0.71808

Table 18.3 Common factors and common variances.

	L_1	L_2	L_3	L_4	h_i^2
x_1	0.93030	-0.22457	-0.18385	-0.16450	0.97675
x_2	0.93619	-0.09199	-0.16230	-0.25179	0.97465
x_3	0.91062	-0.20868	-0.10065	0.21700	0.93000
x_4	0.61632	-0.05391	0.71783	0.29566	0.98546
x_5	0.33628	0.75345	-0.45523	0.32286	0.99225
x_6	0.33018	0.80234	0.37948	-0.30179	0.98785

18.2.2 Iterative principal component method

The principal component method in the above section is the first step in the iterative principal component method, which provides the common variance for each variable, i.e., $h_1^2, h_2^2, \dots, h_p^2$.

Replacing the common variances to the diagonal "1" in the correlation matrix, the new matrix R^* is the following:

$$R^* = \begin{bmatrix} h_1^2 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{12} & h_2^2 & r_{23} & \cdots & r_{2p} \\ * & * & * & & * \\ * & * & * & * & * \\ r_{1p} & r_{2p} & r_{3p} & \cdots & h_p^2 \end{bmatrix}.$$

Using this modified matrix R^* to calculate the eigenvalues and eigenvectors, the m new principal components $C_1^*, C_2^*, \dots, C_m^*$ can be obtained. Using (18.10) and (18.12), we can derive new common factors and factor models. From the new model, we can again derive new common variances $h_1^{*2}, h_2^{*2}, \dots, h_p^{*2}$. Up to now, we have just finished the first iteration.

Continue to repeat the above iteration many times until the updated common factors do not change from the previous iteration.

Iterative principal component method is commonly used in psychology and other social sciences. Many mathematical statisticians use only one-step principal component analysis because it has unambiguous mathematical properties. When original variables follow multivariate normal distributions, we can use the maximum likelihood method to perform a factor analysis. Many software packages have an option to use the maximum likelihood method. However, it is not easy to have multivariate data that follow a multivariate normal distribution.

18.3 Factor Pattern Plot and Factor Rotation

18.3.1 Factor pattern plot

By dropping the individual factor in (18.3), we can approximate the original variables by linear combinations of factors

$$\begin{aligned} x_1 &= L_{11}F_1 + L_{12}F_2 + \cdots + L_{1m}F_m, \\ x_2 &= L_{21}F_1 + L_{22}F_2 + \cdots + L_{2m}F_m, \\ &\vdots \\ x_p &= L_{p1}F_1 + L_{p2}F_2 + \cdots + L_{pm}F_m. \end{aligned}$$

If we treat F_1, F_2, \dots, F_m as orthogonal axes and the loadings of each factor ($L_{i1}, L_{i2}, \dots, L_{im}$) as the coordinates of the corresponding axes, we can plot observed points x_i according to the coordinating system of F_1, F_2, \dots, F_m . The plot of x_1, x_2, \dots, x_p in this system is called factor pattern plot. As we cannot plot more than two dimensions simultaneously, we plot them into pairwise combinations of their common factors. Figure 18.1 shows some two-dimensional factor pattern plots for the results in Table 18.3. We can see that x_1, x_2 and x_3 are close to each other. The pattern of other variables is not clear.

As the closeness of variables can only be expressed in higher dimensions, it is not easy to see it with two-dimensional plots. A possible approach is to use the cluster analysis to calculate distance between x_i and x_j based on their distance in terms of factor loadings.

$$d^2(x_i, x_j) = (L_{i1} - L_{j1})^2 + (L_{i2} - L_{j2})^2 + \dots + (L_{im} - L_{jm})^2.$$

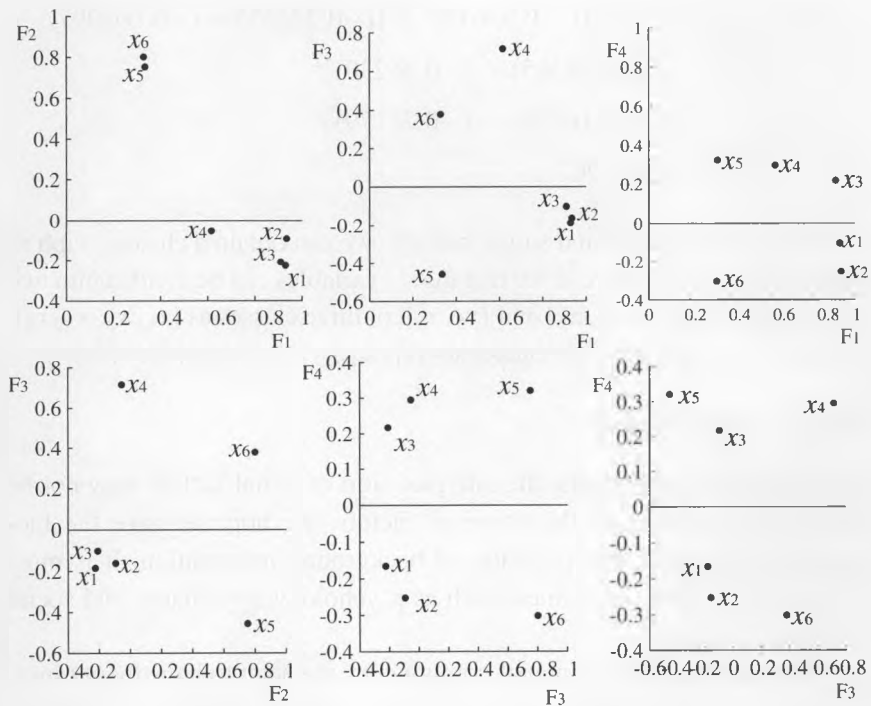


Fig. 18.1 Factor pattern plots for Example 18.1.

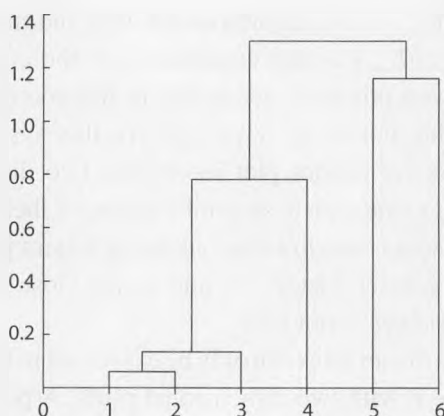


Fig. 18.2 Variable cluster based on common factors.

For example, in Example 18.1, the distance between x_1 and x_2 is

$$\begin{aligned}
 d^2(x_1, x_2) &= (0.93030 - 0.93619)^2 + [(-0.22457) - (-0.09199)]^2 \\
 &\quad + [(-0.18385) - (-0.16230)]^2 \\
 &\quad + [(-0.16450) - (-0.25179)]^2 \\
 &= 0.025696.
 \end{aligned}$$

Using the distances and single linkage, we can obtain a cluster graph as shown in Fig. 18.2. One can see that the six variables can be divided into two main classes, $\{x_1, x_2, x_3, x_4\}$ and $\{x_5, x_6\}$, or three classes as $\{x_1, x_2, x_3, x_4\}$, $\{x_5\}$, and $\{x_6\}$, or even four classes as $\{x_1, x_2, x_3\}$, $\{x_4\}$, $\{x_5\}$ and $\{x_6\}$.

18.3.2 Factor rotation

Like principal components, the interpretation of initial factors may not be clear. Because they are the common factors, we hope to have the factors providing clear interpretation of background information. It is most important for some disciplines, such as psychology, psychiatry, and social sciences.

When we study the principal components, we want to have them represent most of variance of the original variables. In factor analysis, we want the factors to have clear interpretations. However, there is a trade-off between

maximizing the variance and getting clear interpretations. Factor rotation, a procedure to rotate F_1, F_2, \dots, F_m , can be used to achieve both goals.

18.3.2.1 Varimax orthogonal rotation

From Fig. 18.1 or Table 18.3, X_1, X_2 and X_3 have heavy loading on F_1 , but not much on other factors. Can we make F_1 as the factor explaining variation only in body heights? As other variables have also considerable loading on F_1 , we cannot ignore the loadings of chest size, shoulder width, and pelvic width. Similarly, X_5 and X_6 have heavy loadings on F_2 , but they also have considerable loadings on other factors. Therefore, F_2 does not represent all variations in these “sizes”.

To create new factors that have clear interpretations, we want each variable to maximize its loading on one factor and minimize its loading on the other factors. Meanwhile, in order to keep all factors uncorrelated, we have to keep them orthogonal with each other. This can be achieved by orthogonal rotation of the factors, which is called varimax orthogonal rotation. Statistical software packages provide the solutions including the transformation matrix, new factors after rotation, and loadings of variables on the new factors. By rotating the factors in Example 18.1, we get the new factors in Table 18.4 (reading in columns). The variables' loading on the new factors are given in Table 18.5 (reading in rows). Comparing with Table 18.3, the common variance has not changed after the rotation if we ignore small differences due to roundup effect.

Figure 18.3 is a two-dimensional factor plots for Table 18.5. Comparing to Fig. 18.1, the relationship between the variables and the factors became clearer. F'_1 is mainly related to X_1, X_2 and X_3 , and can be called with height

Table 18.4 Coefficients of varimax rotated factors in Example 18.1.

	F'_1	F'_2	F'_3	F'_4
x_1	0.46388	-0.23962	0.06844	-0.17165
x_2	0.48674	-0.34766	0.26338	-0.23174
x_3	0.18967	0.33808	-0.38633	0.29342
x_4	-0.22420	1.03209	-0.06980	0.03326
x_5	-0.12401	0.08623	-0.19294	1.05113
x_6	0.00118	-0.11918	1.00848	-0.17486

Table 18.5 Factor pattern after varimax rotation for Example 18.1.

	F'_1	F'_2	F'_3	F'_4	h_i^2
x_1	0.97579	0.14777	0.02761	0.04439	0.97675
x_2	0.96221	0.10651	0.17923	0.07301	0.97465
x_3	0.82942	0.41639	-0.14046	0.22125	0.93000
x_4	0.27265	0.92843	0.20268	-0.08977	0.98546
x_5	0.13893	-0.06927	0.23452	0.95559	0.99225
x_6	0.04566	0.17267	0.94799	0.23929	0.98785

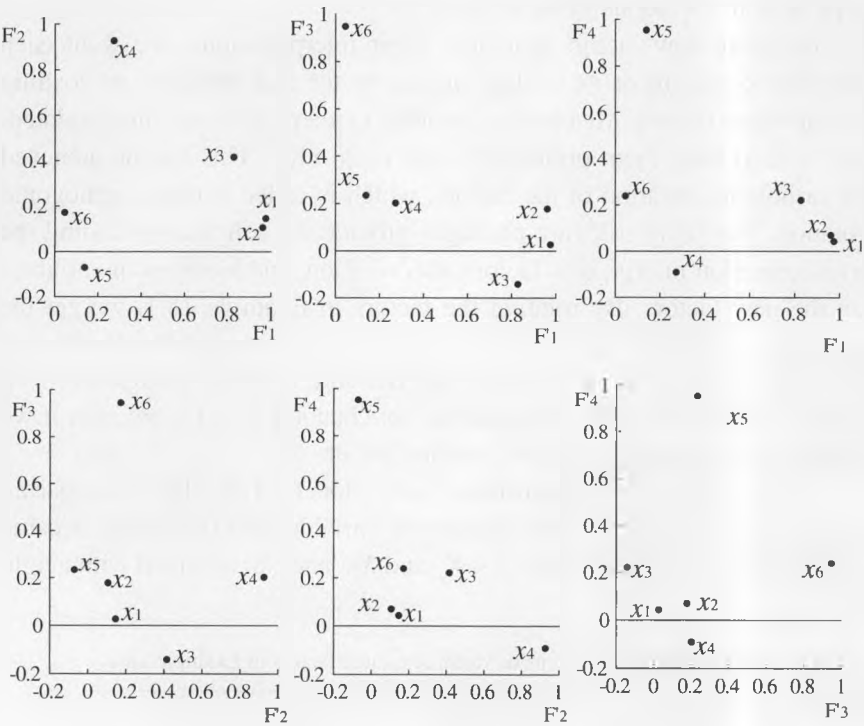


Fig. 18.3 New factor pattern plot of Fig. 18.1 after varimax orthogonal rotation.

factor; F'_2 is mainly related to X_4 , can be called with chest width factor; F'_3 and F'_4 are related to X_6 and X_5 , respectively and can be called pelvic size factor and shoulder width factor. In summary, rotation of the factors can make the new factors more clearly related to their corresponding variables without changing the total common variance.

For illustration purposes, Example 18.1 has only a few variables so that some of the rotated factors only represent one variable, which is not practically meaningful. When the number of variables increases, one factor may represent information of many variables, then the number of factors may significantly reduce. In such cases, we can use only a few factors to perform a regression or discriminant analysis to obtain reliable results.

18.3.2.2 *Oblique rotation*

To make the new factors clearly associate with a few variables only, the non-orthogonal rotation can be a choice, which allows the factors to be correlated with each other. The new factor may go through the cluster of variables in the factor pattern plot, thus, the factor may only represent these variables.

Note that the new factors after non-orthogonal rotation might still have the colinearity problem. Therefore, it is not commonly used, and most people prefer the varimax orthogonal rotation.

18.4 Factor Score and Application of Factor Patterns

18.4.1 *Factor score and principal components*

Factor score is similar to principal component, which reflects the amount of information carried by the factor. If the factors are extracted from the principal components, the factor scores are exactly the scores of standardized principal components before rotation. Through varimax rotation, they no longer maximize the variances along their directions, but they are still orthogonal to each other. The gain of the rotation is to make the interpretation of factors more precise.

Many statistical software packages provide the coefficients to express the final factors as linear combinations of original variables. Using Example 18.4 for an illustration, we can express the factors F'_1 , F'_2 , F'_3

and F_4' in linear combinations of original variables, such as

$$\begin{aligned} F_1' &= 0.47268x_1 + 0.49639x_2 + 0.17444x_3 - 0.23584x_4 \\ &\quad - 0.12558x_5 + 0.00185x_6, \\ F_2' &= -0.24267x_1 - 0.35181x_2 + 0.36635x_3 + 1.00877x_4 \\ &\quad + 0.09411x_5 - 0.14130x_6. \end{aligned}$$

Here, x_1, x_2, \dots, x_6 are the standardized variables.

We can use the software package to calculate the factor scores for each individual and use them in further statistical analysis. Therefore, factor analysis is an extension of principal component analysis, which can achieve the goals of the principal component analysis in terms of summarizing the variables into a few comprehensive scores.

18.4.2 *Common factors and latent variables*

When we treat the common factors F_1, F_2, \dots, F_m as latent variables that we cannot directly observe, the factor scores based on the observed values of X_1, X_2, \dots, X_p can improve our understanding of the underlying biological processes. This is part of the reason that we like the factors to have clear interpretations.

18.4.3 *Factor analysis and structure validity*

Suppose there are two sets of questionnaires A and B . A is an existing acceptable instrument, while B is newly developed. To illustrate that the questionnaire B measures the same underlying structure as A , we may collect the measurements by both questionnaires simultaneously from a proper sample; then perform factor analysis for both sets respectively and compare their factor structures, including their variances and interpretations. If the common factors of A and B have similar structures, we can conclude the structure equivalence between B and A .

18.5 *Confirmatory Factor Analysis*

The previous introduced methods are subject to exploratory factor analysis (EFA) because we do not have prior knowledge of the factor structure

of the original variables, do not know the number of factors, the specific relationship between factors and variables. The answer is only available after the analysis.

Confirmatory factor analysis (CFA) is different from exploratory factor analysis. We use the following example to introduce the CFA and explain the differences between CFA and EFA.

Example 18.2 (Byrne (1994)) Maslach Burnout Inventory (MBI) is a psychological measure for “burnout.” The term “burnout” is related to a situation increasingly arising with stress levels that impairs people’s normal function level. According to its design, MBI measures decreases in emotional exhaustion (EE), depersonalization (DP), personal accomplishments (PA). EE measures feelings of being emotionally over extended and exhausted by one’s work. DP measures an unfeeling and impersonal response toward recipients of one’s service, care treatment or instruction. PA measures feelings of competence and successful achievement in one’s work. There are a total of 22 questions, 9 for EE, 5 for DP, and 8 for PA. (From Byrne (1994).)

In its design, research prospectively assumed that the questions to be asked reflected the latent factor structure of EE, DP and PA. However, one did not know if the questionnaire could achieve the expected goal in practice, i.e., if the answers of questions could be attributed into these three factors. 372 male middle school teachers were invited to answer the MBI questions. The structure validity of the questionnaire needed to be confirmed based on this dataset of real responses.

First, according to the design of MBI, the factor pattern plot should be similar to Fig. 18.4. This figure assumed that (1) there should be three factors; (2) each question should have nonzero loadings for the corresponding factors; (3) the three factors are related (with bi-direction arrows); (4) the responses to different questions should be independent (with only one arrow from a factor).

Meanwhile, the author proposed an alternative structure of factor in Fig. 18.5. In the new model, question 12 should have nonzero loadings in both PA and EE factors. In addition, the responses to questions 1 and 2, 6 and 16, as well as 10 and 11 were correlated (see the bi-direction arrows).

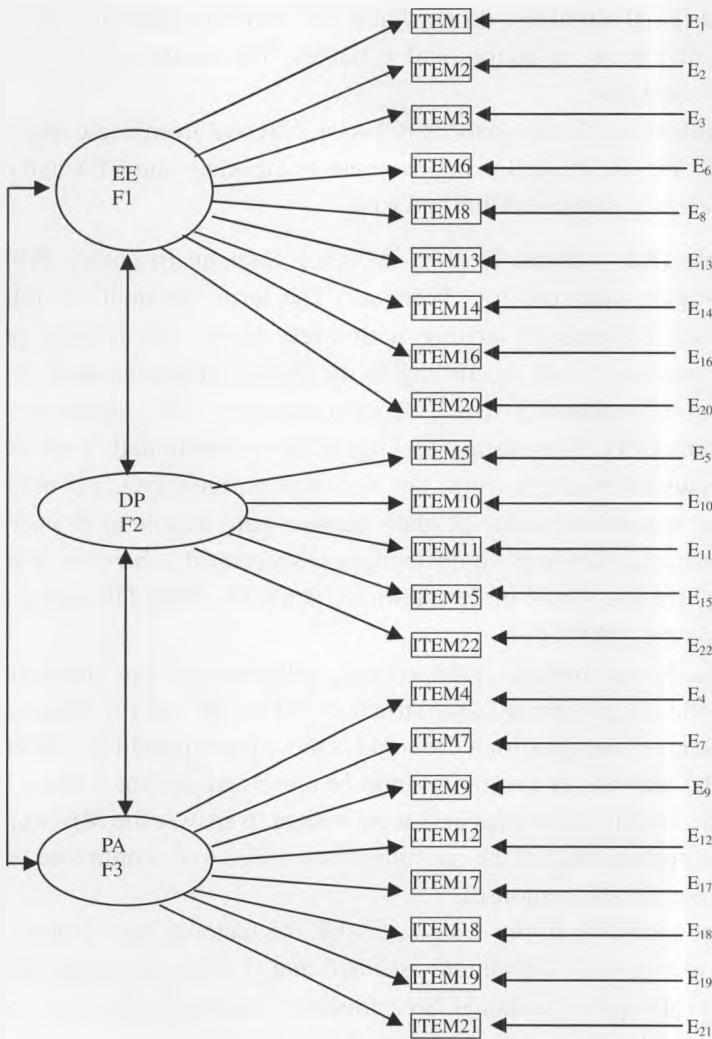


Fig. 18.4 Factor pattern plot 1 of MBI.

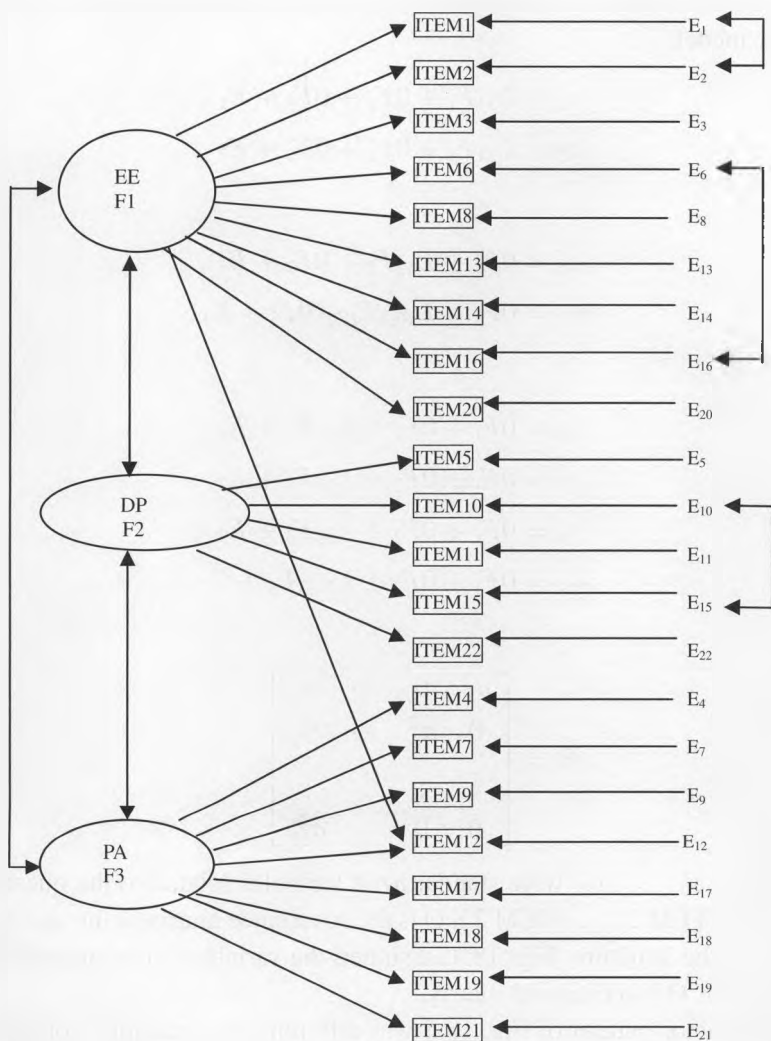


Fig. 18.5 Factor pattern plot 2 of MBI.

Using the factor structure in Fig. 18.4, one could derive the following factor model:

$$\begin{aligned}
 x_1 &= L_{11}F_1 + 0F_2 + 0F_3 + E_1 \\
 x_2 &= L_{21}F_1 + 0F_2 + 0F_3 + E_2 \\
 &\vdots \\
 x_5 &= 0F_1 + L_{52}F_2 + 0F_3 + E_5 \\
 x_{10} &= 0F_1 + L_{10,2}F_2 + 0F_3 + E_{10} \\
 &\vdots \\
 x_4 &= 0F_1 + 0F_2 + L_{43}F_3 + E_4 \\
 x_7 &= 0F_1 + 0F_2 + L_{73}F_3 + E_7 \\
 x_9 &= 0F_1 + 0F_2 + L_{93}F_3 + E_9 \\
 x_{12} &= 0F_1 + 0F_2 + L_{12,3}F_3 + E_{12} \\
 &\vdots
 \end{aligned}$$

$$\Theta = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{22}^2 \end{bmatrix}$$

where x_1, x_2, \dots, x_{22} were standardized variables related to the questions ITEM 1, ITEM 2, \dots , ITEM 22. Θ is the covariance matrix of the variables. Because the structure, Fig. 18.4, assumed the variables to be uncorrelated each other, Θ is a diagonal matrix.

The factor pattern of Fig. 18.5 was different. For example, comparing to the above equations, the equation for X_{12} should be changed to

$$x_{12} = L_{12,1}F_1 + 0F_2 + L_{12,3}F_3 + E_{12}$$

and the covariance matrix Θ was no longer a diagonal matrix. Instead, the covariance at the first column and second row should be $\sigma_{1,2}$ and that at the sixth column and 16th row should be $\sigma_{6,16}$.

After the factor models were specified, one could perform a goodness-of-fit test to compare the two models with the observed data. The

computational processes of goodness-of-fit test is complicated and beyond the intended level of this book. There are specialized software packages to perform such analysis, such as LISREL and EQS. The indices for goodness-of-fit are χ^2 statistics and goodness-of-fit index (GFI). χ^2 statistics is sensitive to the sample size and deviations from the assumption of normal distribution. Therefore, it is suggested to use χ^2 statistics as a measure rather than a test statistics for how good the model fits the data. When the degrees of freedom are the same, the model with a larger value of the χ^2 statistics fits data better than the model with a smaller value of the χ^2 statistics. Usually, we take difference of two χ^2 statistics when the degrees of freedom are different. The GFI is the value of change in χ^2 statistics relative to change in degrees of freedom and it takes values between 0 and 1. Usually, GFI should be above 0.9 to indicate a reasonable model. GFI is not a function of sample size, which is more robust when the distributions deviate from multivariate normal distributions. Thus, it is often used to measure the goodness-of-fit.

As shown in Table 18.6, in our example, for the model in Fig. 18.4, the goodness-of-fit was measured by a χ^2 statistics of 693.849 with degrees of freedom 206, and $\text{GFI} = 0.848$. According to the standard by experience (0.9 or larger), the model could be considered unsatisfied. For the model in Fig. 18.5, the χ^2 statistics reduced to 445.219 and degrees of freedom was 202, a reduction of 4 from Model 1. The GFI was 0.924. Thus, Model 2 fitted the data better.

Example 18.2 demonstrated the difference between EFA and CFA. On one hand, any variable could have loadings on all possible factors in EFA, and the loadings become known only after the analysis. On the other hand, variables in CFA could have loadings only on the assumed factors and the loadings on other factors are set to zero before analysis.

Table 18.6 Results of goodness-of-fit for two factor models.

Models	χ^2 statistics	Degrees of freedom	GFI
Model 1	693.849	206	0.848
Model 2	445.219	202	0.924

Detailed theory and methods of confirmatory factor analysis can be found in books of Byrne (1994) and Ke *et al.* (1992).

18.6 Computerized Experiments

Experiment 18.1 Calculations on exploratory factor analysis Using data of 1985 Han male young adults (aged 19–22) from 28 Chinese provinces/cities to calculate all the latent factors.

Lines 01 to 06 of Program 18.1 read data into SAS dataset A. Line 05 is not SAS language, it is a short notation to skip listing of all records. Line 07 asks program to calculate correlation matrix. Lines 08 to 12 perform a factor analysis. The option METHOD=P specifies the use of principal component analysis. Other choices are PRINIT for iterative principal factor analysis and ML for the maximum likelihood factor analysis. Option ROTATE specifies the rotation method with choices of V for varimax method and PROCRUSTES for non-orthogonal rotations. PREPLOT

Program 18.1 Factor analysis.

Line	Program
01	DATA A;
02	INPUT X1-X6;
03	CARDS;
04	173.28 93.62 60.10 86.72 38.97 27.51

	168.99 91.52 55.11 86.23 38.30 27.14
05	;
06	PROC CORR DATA=A OUT=CORREL;
07	PROC FACTOR DATA=CORREL PREPLOT SCREE METHOD=P ROTATE=V PLOT;
08	PROC FACTOR DATA=CORREL PREPLOT SCREE METHOD=PRINIT PRIORS=R ROTATE=V PLOT;
09	PROC FACTOR DATA=CORREL PREPLOT SCREE METHOD=ML HEYWOOD ROTATE=V PLOT;
10	PROC FACTOR DATA=CORREL PREPLOT SCREE METHOD=P ROTATE=PROCRUSTES;
11	PROC FACTOR DATA=CORREL OUTSTAT=FACT METHOD=P ROTATE=V SCORE;
12	PROC SCORE DATA=A SCORE=FACT OUT=SCORES;
13	PROC PRINT DATA=SCORES;
14	RUN;

asks to plot factor pattern plots before they are rotated. PLOT asks for factor pattern plots after rotations. Lines 13 and 14 calculate the factor scores. Factor scores can also be derived by changing option on line 13 OUTSTAT=FACTOR into OUT=SCORES. In this case, line 14 can be skipped.

Experiment 18.2 Principal components and common factors Use data of Experiment 17.1 to perform an exploratory factor analysis; observe and discuss the factor model and factor pattern plot; examine the possible relationship between the common factors and the principal components.

Experiment 18.3 Factor score regression and principal component regression Using data of Experiment 17.2 to perform an exploratory factor analysis and regression analysis of factors to obtain factor scores. Compare your results with the results of Experiment 17.3.

Readers can use SAS programs in Chap. 17 and SAS Program 18.1 to develop their own programs for these two exercises.

18.7 Practice and Experiments

1. Using the 1985 Chinese young Han adults (19–22 years old) functional data in the exercises of Chap. 17 to perform an exploratory factor analysis.
2. Combining both functional and body development data of 1985 Chinese young Han adults (19–22 years old) from 28 Chinese provinces/cities to perform a factor analysis. Discuss your results, in particular, the relationship between development variables and functional variables.
3. Perform a cluster analysis based on the factors derived from the above problem, and compare with the results of factor plot based on the factor analysis. Comment on their similarities and differences and explore the possible reasons.
4. For a dependent variable Y and p independent variables X_1, X_2, \dots, X_p , should we use a factor analysis to perform stepwise selection of independent variables? Use the data in Chap. 23 to compare the variables selected based on the use of multiple regression versus the factor analysis.

5. Suppose that you have observed a sample of variables X_1, X_2, \dots, X_p and you already have the coefficients (eigenvectors) and variances (eigenvalues) of p principal components. Can you use a calculator to derive the common factors F_1, F_2, \dots, F_p and their corresponding factor loadings for all variables? Use your answer to discuss the relationship between factor analysis and principal component analysis.

(1st edn. Ying Lu; 2nd edn. Yuanto Hao, Lifen Feng)

Chapter 19

Canonical Correlation and Correspondence Analysis

In Chap. 9, the simple correlation between two random variables Y and X is introduced. The multiple correlation between a random variable Y and a set of random variables X_1, X_2, \dots, X_p is discussed in Chap. 12. In this chapter, we will discuss canonical correlation between two sets of random variables, Y_1, Y_2, \dots, Y_q and X_1, X_2, \dots, X_p , as well as the correspondence analysis of contingency tables.

19.1 Canonical Correlation

In medical research, we often need to analyze the correlation between one set of p indices and another set of q indices. A simple approach is to calculate pq pairwise correlation coefficients, to describe correlations between the two sets of variables. This approach however is somewhat tedious, and more importantly, does not solve the main problem. A more efficient approach is to consider, as in principal component analysis, linear combinations of variables in each set. From these two sets of linear combinations, we look for the “most” correlated combinations. Thus the correlation between the two sets of variables can be described by two comprehensive scores, which reveals the best concise relationship. This is the main idea behind the canonical correlation analysis. Furthermore, similar to the relationship between simple correlation and simple regression, and between multiple correlation and multiple regression, canonical correlation can also lead to the regression of one set of variables on another set of variables.

Example 19.1 Correlation between morphology indices and function indices Let us go back to the 1985 survey of physical data of male students

from 28 Chinese cities. Let X_1, X_2, \dots, X_6 denote the morphology indices, and Y_1, Y_2, \dots, Y_5 denote the function indices (see Chap. 17 for the original data). We wish to investigate the correlation between the two sets of variables.

Prior to any multivariate analysis, we usually summarize the data by descriptive statistics like mean, standard deviation, and correlation coefficients. Varying from one to another, each variable has a specific measurement unit and its range. Therefore, the first step is usually a standardization of the variables.

A first description of the correlation between morphology and function indices is the correlation matrix Table 19.1. Here, the upper right block is just the correlation between X_1, X_2, \dots, X_6 and Y_1, Y_2, \dots, Y_5 . Under-scored values are tested significant correlation coefficients. Theoretically, only when this block of correlation coefficients is all non-significant, can we say X_1, X_2, \dots, X_6 and Y_1, Y_2, \dots, Y_5 are not correlated. However, it is sometimes too cumbersome to describe the correlation between two sets of variables with a correlation matrix. Moreover, it is hard to assess the impact of one significant correlation coefficient on the correlation of two sets of variables. A more comprehensive measure of correlation is thus desired.

If we partition the correlation matrix of $(X_1, X_2, \dots, X_p, Y_1, Y_2, \dots, Y_q)$ into three submatrices, upper left, upper right, and lower right, correspondingly denoted by R_{XX} , R_{XY} and R_{YY} , then the above matrix can be expressed as follows:

$$R = \begin{pmatrix} R_{XX} & R_{XY} \\ R_{XY} & R_{YY} \end{pmatrix}, \quad (19.1)$$

where R_{XX} and R_{YY} represent the correlation matrices of variables X and variables Y respectively, and R_{XY} is the correlation matrix between X and Y .

19.1.1 Definition of canonical correlation variables

Suppose there are independent variables X_1, X_2, \dots, X_p and dependent variables Y_1, Y_2, \dots, Y_q ($p = 6, q = 5$ as in the example).

(1) Standardize all the variables, and denote the standardized variables by x_1, x_2, \dots, x_p and y_1, y_2, \dots, y_q .

(2) Identify the first pair of canonical variables by finding the appropriate coefficients $(a_{11}, a_{12}, \dots, a_{1q})$ and $(b_{11}, b_{12}, \dots, b_{1p})$ such that

$$\begin{aligned} W_1 &= a_{11}y_1 + a_{12}y_2 + \dots + a_{1q}y_q, \\ V_1 &= b_{11}x_1 + b_{12}x_2 + \dots + b_{1p}x_p, \end{aligned} \quad (19.2)$$

and their correlation coefficient $\text{Corr}(W_1, V_1)$ is maximized.

W_1 and V_1 are called the first pair of canonical correlation variables, or simply canonical variables. $\text{Corr}(V_1, W_1)$ is called the first canonical correlation coefficient, denoted by $R_{1, \text{Can}}$.

From Chap. 9 we know that linear transformation does not change the correlation among variables. Therefore, there are infinitely many linear coefficients that satisfy the above conditions. To make the coefficients $(a_{11}, a_{12}, \dots, a_{1q})$ and $(b_{11}, b_{12}, \dots, b_{1p})$ unique, we put a further restriction

$$\mathbf{a}'_1 \mathbf{R}_{XX} \mathbf{a}_1 = 1, \quad \mathbf{b}'_1 \mathbf{R}_{YY} \mathbf{b}_1 = 1. \quad (19.3)$$

Here, vectors $\mathbf{a}' = (a_{11}, a_{12}, \dots, a_{1q})$, $\mathbf{b}' = (b_{11}, b_{12}, \dots, b_{1p})$. Thus, the aforementioned process is to maximize $R_{1, \text{Can}} = \text{Corr}(W_1, V_1)$ under the conditions of $\mathbf{a}'_1 \mathbf{R}_{XX} \mathbf{a}_1 = 1$, $\mathbf{b}'_1 \mathbf{R}_{YY} \mathbf{b}_1 = 1$.

(3) Identify the second pair of canonical variables. First we find a second set of linear coefficients $(a_{21}, a_{22}, \dots, a_{2q})$ and $(b_{21}, b_{22}, \dots, b_{2p})$, such that

$$\begin{aligned} W_2 &= a_{21}y_1 + a_{22}y_2 + \dots + a_{2q}y_q, \\ V_2 &= b_{21}x_1 + b_{22}x_2 + \dots + b_{2p}x_p. \end{aligned}$$

And

- (i) W_2 is uncorrelated with W_1 and V_1 ;
- (ii) V_2 is uncorrelated with W_1 and V_1 ;
- (iii) $R_{2, \text{Can}} = \text{Corr}(W_2, V_2)$ is maximized under the conditions of $\mathbf{a}'_2 \mathbf{R}_{XX} \mathbf{a}_2 = 1$, $\mathbf{b}'_2 \mathbf{R}_{YY} \mathbf{b}_2 = 1$.

(4) Identify the k th pair of canonical variables. As we proceed above, W_k, V_k are linear combinations of y_1, y_2, \dots, y_q and x_1, x_2, \dots, x_p respectively, which are uncorrelated with canonical variables $W_1, V_1, W_2, V_2, \dots, W_{k-1}, V_{k-1}$, and $R_{k, \text{Can}} = \text{Corr}(W_k, V_k)$ is maximized

under the conditions of

$$\mathbf{a}_k' \mathbf{R}_{XX} \mathbf{a}_k = 1, \mathbf{b}_k' \mathbf{R}_{YY} \mathbf{b}_k = 1.$$

Continuing this process, we can identify at most s pairs of canonical variables, $s = \min(p, q)$. Obviously,

$$\text{Corr}(W_1, V_1) \geq \text{Corr}(W_2, V_2) \geq \cdots \geq \text{Corr}(W_s, V_s). \quad (19.4)$$

The calculations above are all based on the correlation coefficient matrix. Therefore, strictly speaking, the canonical variables thus obtained should be called standardized canonical variables. If the calculations are based on variance-covariance matrix, the canonical variables will be called canonical variables, without the modifier "standardized". Note that the canonical variables obtained based on the two matrices are different in general. Most authors prefer using the correlation matrix to obtain standardized canonical variables. Therefore the adjective "standardized" is usually omitted.

19.1.2 *Calculations of canonical correlation coefficients and linear combination coefficients*

In actual computations, correlation coefficient matrix (19.1) of $(X_1, X_2, \dots, X_p, Y_1, Y_2, \dots, Y_q)$ is obtained first. Finding canonical correlation coefficients reduces to the singular value decomposition of matrix $\mathbf{R}_{XX}^{-1} \mathbf{R}_{XY} \mathbf{R}_{YY}^{-1}$, which is to be expressed as,

$$\mathbf{R}_{XX}^{-1} \mathbf{R}_{XY} \mathbf{R}_{YY}^{-1} = \mathbf{V} \mathbf{S} \mathbf{W} \quad (19.5)$$

where \mathbf{S} is a diagonal matrix of eigenvalues, and \mathbf{V} and \mathbf{W} are orthogonal matrices of the column eigenvectors. From the i th eigenvalue λ_i , we can obtain the following i th canonical correlation coefficient $R_{i,Can}$,

$$R_{i,Can} = \sqrt{\frac{\lambda_i}{1 + \lambda_i}}.$$

Its approximate standard error is

$$SE[R_{i,Can}] = \frac{1 - R_{i,Can}^2}{\sqrt{n-1}} = \frac{1}{(1 + \lambda_i)\sqrt{n-1}}.$$

The eigenvectors are just the linear combination coefficients. Furthermore, the magnitude of each eigenvalue as a percentage of the sum of

eigenvalues represents the percentage of correlation information of each canonical variable.

By definition, all canonical correlation coefficients are between 0 and 1. Since the canonical correlation coefficients measure the correlation between two sets of variables to the maximum, the first canonical correlation coefficient is no less than the absolute value of any individual correlation coefficient, i.e.,

$$R_{1,Can} \geq \max(|\text{Corr}(x_i, y_j)|).$$

From the data in the example above, we can obtain Table 19.2 for eigenvalues, percentage of information, and canonical correlation coefficients, and Table 19.3 for linear combination coefficients. From Table 19.2, one

Table 19.2 Eigenvalues and canonical correlation coefficients.

	Eigenvalue	Percentage	Cumulative percentage	Canonical correlation coefficient	Approximate standard error
1	3.2422	0.6546	0.6546	0.874228	0.045366
2	1.1912	0.2405	0.8951	0.737312	0.087829
3	0.3533	0.0713	0.9665	0.510941	0.142209
4	0.1436	0.0290	0.9955	0.354369	0.168283
5	0.0235	0.0045	1.0000	0.148207	0.188223

Table 19.3 Linear combination coefficients for standardized canonical correlations.

	V_1	V_2	V_3	V_4	V_5
X_1	0.5852	-1.1443	0.7823	0.0352	-0.8298
X_2	-0.2175	0.0189	0.6032	0.1289	1.5590
X_3	0.5288	1.6213	-0.7370	-0.4066	-1.1704
X_4	0.1890	-0.9874	-0.7753	0.1229	0.6988
X_5	-0.1193	-0.0626	-0.2509	-0.5860	1.0488
X_6	0.1948	0.8108	0.1467	0.9523	-0.5140
	W_1	W_2	W_3	W_4	W_5
Y_1	-0.0838	-0.1325	1.0807	0.3750	-0.0376
Y_2	-0.0878	1.2688	0.0701	0.2476	-0.3342
Y_3	0.2147	-0.3301	0.2218	-1.0863	1.4100
Y_4	0.2920	-0.2392	-0.5765	1.3368	-0.2942
Y_5	0.7607	-0.2995	0.6532	-0.0017	-0.6905

can conclude that first two pairs of canonical variables contain about 90% of all correlation information, together with the third pair, they contain 96.65% of correlation information. The first two canonical correlation coefficients are relatively high at 0.8743 and 0.7373, and even the third canonical correlation coefficient has reached 0.5109.

The first canonical variables are

$$W_1 = -0.0838y_1 - 0.0878y_2 + 0.2147y_3 + 0.2920y_4 + 0.7607y_5,$$

$$V_1 = 0.5852x_1 - 0.2175x_2 + 0.5288x_3 + 0.1890x_4 - 0.1193x_5 \\ + 0.1948x_6.$$

The second canonical variables are

$$W_2 = -0.1325y_1 + 1.2688y_2 - 0.3301y_3 - 0.2392y_4 - 0.2995y_5,$$

$$V_2 = -1.1443x_1 - 0.0189x_2 + 1.6213x_3 - 0.9874x_4 - 0.0626x_5 \\ + 0.8108x_6.$$

19.1.3 Canonical correlation structures

The canonical correlation coefficients quantitatively describe the linear correlation between two sets of variables. We now turn to the relationship between the canonical variables and their original variables. In fact, this relationship has much to do with the original correlation coefficients between the variables and the linear combination coefficients. Using software packages, it is not hard to calculate correlation coefficients between the variables and their corresponding canonical variables, and the correlation coefficients between the variables and the canonical variables of the other set.

It is worth noting that the correlation coefficient between an original variable and a canonical variable of the other set is the product of the correlation coefficient of the variable with its own canonical variable and the canonical correlation coefficient. Formally,

$$\text{Corr}(x_i, W_j) = R_{j,\text{Can}} \text{Corr}(x_i, V_j),$$

$$\text{Corr}(y_i, V_j) = R_{j,\text{Can}} \text{Corr}(y_i, W_j).$$

Table 19.4 displays the correlation coefficients between the original variables and their canonical variables (underscored values are greater than 0.5).

Table 19.4 Loadings of original variables on canonical variables (i.e., correlation coefficient between the original variables and the canonical variables).

	V_1	V_2	V_3	V_4	V_5	W_1	W_2	W_3	W_4	W_5
X_1	<u>0.9050</u>	-0.0806	0.3777	-0.1487	0.0887	<u>0.7912</u>	-0.0594	0.1930	-0.0527	0.0132
X_2	<u>0.8616</u>	0.0112	0.4152	-0.0360	0.2412	<u>0.7532</u>	0.0083	0.2121	-0.0128	0.0357
X_3	<u>0.9361</u>	0.1655	-0.0471	-0.2933	-0.0247	<u>0.8184</u>	0.1220	-0.0240	-0.1039	-0.0037
X_4	<u>0.6958</u>	-0.3189	<u>-0.5382</u>	0.3191	0.1354	<u>0.6083</u>	-0.2351	-0.2750	0.1131	0.0201
X_5	0.1356	<u>0.5329</u>	-0.0321	-0.2376	<u>0.7389</u>	0.1185	0.3929	-0.0164	-0.0842	0.1095
X_6	0.2433	0.4412	-0.0405	<u>0.7478</u>	0.3908	0.2127	0.3253	-0.0207	0.2650	0.0579
Y_1	-0.3610	-0.0625	0.3757	0.1605	0.0410	-0.4130	-0.0848	<u>0.7353</u>	0.4530	0.2764
Y_2	0.3963	<u>0.6232</u>	0.0495	0.0508	0.0332	0.4533	<u>0.8452</u>	0.0968	0.1433	0.2240
Y_3	<u>0.5801</u>	0.1568	0.0378	0.0287	0.1050	<u>0.6636</u>	0.2127	0.0740	0.0810	<u>0.7087</u>
Y_4	<u>0.5003</u>	0.0296	-0.0837	0.2339	0.0677	<u>0.5723</u>	0.0401	-0.1638	<u>0.6600</u>	0.4565
Y_5	<u>0.7994</u>	0.0094	0.0685	-0.0743	-0.0473	<u>0.9144</u>	0.0128	0.1341	-0.2098	-0.3190

These correlations are referred to as the loadings of the original variables on the canonical variables. The upper left and lower right blocks represent the correlation among original variables and their corresponding canonical variables, providing basis for explanation of the canonical variables; the upper right and lower left blocks are the correlation between the original variables and the canonical variables of the other set of variables, providing basis for prediction (regression) of the original variables from the other set.

The upper left block in Table 19.4 indicates, for example,

$$x_1 = 0.9050V_1 - 0.0806V_2 + 0.3777V_3 - 0.1487V_4 + 0.0887V_5,$$

$$x_2 = 0.8616V_1 + 0.0112V_2 + 0.4152V_3 - 0.0360V_4 + 0.2412V_5.$$

From these loadings, we can see that canonical variable V_1 mainly represents the variables of standing height X_1 , sitting height X_2 , body weight X_3 , and chest circumference X_4 ; while V_2 mostly represents the variable of shoulder width X_5 ; V_3 represents chest circumference X_4 ; V_4 represents pelvic diameter X_6 ; and V_5 also represents shoulder width X_5 .

The lower right block in Table 19.4 indicates, for example

$$y_1 = -0.4130W_1 - 0.0848W_2 + 0.7353W_3 + 0.4530W_4 + 0.2764W_5,$$

$$y_2 = 0.4533W_1 + 0.8452W_2 + 0.0968W_3 + 0.1433W_4 + 0.2240W_5.$$

Also we can see that the canonical variable W_1 mostly reflects the variables of diastolic pressures Y_3 , Y_4 and lung capacity Y_5 ; W_2 mainly represents systolic pressure Y_2 ; W_3 represents pulse Y_1 ; W_4 mostly represents diastolic pressure (vanishing sound) Y_4 ; W_5 represents diastolic pressure (changing pitch) Y_3 .

The relationships among the canonical variables are clearly revealed by the apparent representation of individual variables in Table 19.5 below. For instance, the first canonical correlation coefficient 0.8742 reflects the positive correlation between standing height, sitting height, body weight, chest circumference and diastolic pressure, lung capacity, meaning that the larger body frame implies higher diastolic pressure and larger lung capacity. Also, the third canonical correlation coefficient 0.5109 mainly reflects the negative correlation between chest circumference and pulse, i.e., the smaller the chest size, the faster the pulse rate. In general, Table 19.5 agrees well with the physiological principles.

Table 19.5 Interpretation of canonical variables.

	V	W	Corr(V, W)
1	standing height, sitting height, body weight, chest circumference	diastolic pressure, lung capacity	0.8742
2	shoulder width	systolic pressure	0.7373
3	chest circumference (-)	pulse	0.5109
4	pelvic diameter	diastolic pressure (sounds disappear)	0.3544
5	shoulder width	diastolic pressure (sounds muffled)	0.1482

19.1.4 *Rotation of canonical variables and adjusted canonical correlation coefficient*

Similar to the cases in factor analysis, if the canonical variables $V_1, V_2, \dots, W_1, W_2, \dots$ do not provide practical interpretation, we may choose to sacrifice the maximization requirement of $\text{Corr}(W_1, V_1), \text{Corr}(W_2, V_2), \dots$ etc., by obtaining more meaningful new variables $V'_1, V'_2, \dots, W'_1, W'_2, \dots$ via orthogonal rotations. The new correlation coefficients $\text{Corr}(W'_1, V'_1), \text{Corr}(W'_2, V'_2)$ etc. are called adjusted canonical correlation. In the above example, the adjusted correlation coefficients are 0.8256, 0.6634, 0.3728, 0.2854 and 0.0970.

19.1.5 *Redundancy of canonical variables*

Just as the total variance remains unchanged in factor analysis, in canonical analysis, the total variance also remains unchanged if the number of canonical variables is the same as that of the original variables. In the above example, for instance, the number of original variables Y_1, Y_2, \dots, Y_5 is the same as the number of canonical variables W_1, W_2, \dots, W_5 . Therefore,

$$\begin{aligned} \text{Var}(W_1) + \text{Var}(W_2) + \dots + \text{Var}(W_5) \\ = \text{Var}(Y_1) + \text{Var}(Y_2) + \dots + \text{Var}(Y_5) = 5. \end{aligned} \quad (19.6)$$

However, if the number of canonical variables is less than that of the original variables, the total variance will be somewhat reduced. In the example above, the number of canonical variables V_1, V_2, \dots, V_5 is less than the

Table 19.6 Percentage of the variances explained by the canonical variables.

Canonical variable	$X_1, X_2, X_3, X_4, X_5, X_6$				
	Explained by V_1, V_2, \dots, V_5		Canonical correlation coefficient	Explained by W_1, W_2, \dots, W_5	
	Percentage	Cumulative percentage		Percentage	Cumulative percentage
1	0.4999	0.4999	0.7643	0.3821	0.3821
2	0.1024	0.6023	0.5436	0.0557	0.4377
3	0.1016	0.7039	0.2611	0.0265	0.4643
4	0.1378	0.8417	0.1256	0.0173	0.4816
5	0.1306	0.9724	0.0220	0.0029	0.4844

	Y_1, Y_2, Y_3, Y_4, Y_5				
	Explained by W_1, W_2, \dots, W_5			Explained by V_1, V_2, \dots, V_5	
	Percentage	Cumulative percentage		Percentage	Cumulative percentage
1	0.3960	0.3960	0.7643	0.3027	0.3027
2	0.1537	0.5497	0.5436	0.0836	0.3863
3	0.1201	0.6698	0.2611	0.0313	0.4176
4	0.1424	0.8122	0.1256	0.0179	0.4355
5	0.1878	1.0000	0.0220	0.0041	0.4396

number of variables X_1, X_2, \dots, X_6 , therefore,

$$\begin{aligned} & \text{Var}(V_1) + \text{Var}(V_2) + \dots + \text{Var}(V_5) \\ & < \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_6) = 6. \end{aligned} \quad (19.7)$$

The percentage of variance explained by a canonical variable is defined as the variance of the canonical variable as a percentage of the total number of the original variables. Table 19.6 displays the percentages of variances of the original variables explained by the canonical variables. This is so-called redundancy analysis. The percentages of variances explained by the original variables' own canonical variables are $\text{Var}(V_i)/p$ and $\text{Var}(W_j)/q$ respectively. The percentages explained by the other canonical variables are the products of $\text{Var}(V_i)/p$, or $\text{Var}(W_j)/q$, with the squared canonical correlation coefficients. Hence, V_1, V_2, \dots, V_5 do not quite contain all the information of the X variables. Whereas, W_1, W_2, \dots, W_5 contain all the information of the Y variables. Only 48.44% of the information of X variables is contained in W_1, W_2, \dots, W_5 ; and only 43.96% of the information of Y variables is contained in V_1, V_2, \dots, V_5 .

19.1.6 Regression on canonical variables

Back to Table 19.4, the upper right block is the matrix of correlation coefficients between X variables and W_1, W_2, \dots . If we perform a multiple regression of X_i on W_1, W_2, \dots , the determination coefficient will be the sum of corresponding squared correlation coefficients. For example,

$$X_1 \sim W_1 \qquad R^2 = (0.7912)^2 = 0.6260$$
$$X_1 \sim W_1, W_2 \qquad R^2 = (0.7912)^2 + (-0.0594)^2 = 0.6296$$
$$X_1 \sim W_1, W_2, W_3 \qquad R^2 = (0.7912)^2 + (-0.0594)^2 + (0.1930)^2$$
$$\qquad\qquad\qquad = 0.6668$$

etc. Correspondingly, the lower left block of Table 19.4 is the matrix of correlation coefficients between Y variables and V_1, V_2, \dots . Similarly, if we perform a multiple regression of Y_j on V_1, V_2, \dots , the determination coefficient will be the sum of corresponding squared correlation coefficients. Table 19.7 gives the determination coefficients of the regressions of the original variables on the first m canonical variables of the other set of variables.

19.2 Correspondence Analysis

Correspondence analysis was first proposed by French mathematician J.P. Beozecri in 1970. It is primarily used to analyze the correspondence

Table 19.7 The determination coefficients using canonical variables of another set to explain the original variables.

	W_1	$W_1 W_2$	$W_1 W_2 W_3$	$W_1 W_2 W_3 W_4$	$W_1 W_2 W_3 W_4 W_5$
X_1	0.6260	0.6296	0.6668	0.6696	0.6697
X_2	0.5674	0.5674	0.6124	0.6126	0.6139
X_3	0.6697	0.6846	0.6852	0.6960	0.6960
X_4	0.3701	0.4253	0.5010	0.5138	0.5142
X_5	0.0141	0.1684	0.1687	0.1758	0.1878
X_6	0.0452	0.1511	0.1515	0.2217	0.2251
	V_1	$V_1 V_2$	$V_1 V_2 V_3$	$V_1 V_2 V_3 V_4$	$V_1 V_2 V_3 V_4 V_5$
Y_1	0.1303	0.1342	0.2754	0.3012	0.3028
Y_2	0.1571	0.5454	0.5479	0.5505	0.5516
Y_3	0.3366	0.3612	0.3626	0.3634	0.3745
Y_4	0.2503	0.2512	0.2582	0.3129	0.3175
Y_5	0.6390	0.6391	0.6438	0.6493	0.6516

Table 19.8 Summary of the data on the color eyes and that of the hair for 5387 pupils.

Color of the eyes	Color of the hair					Total
	Fair	Red	Medium	Dark	Black	
Dark	98	48	403	681	85	1315
Medium	343	84	909	412	26	1774
Blue	326	38	241	110	3	718
Light	688	116	584	188	4	1580
Total	1455	286	2137	1391	118	5387

Source: Michael J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, 1984, pp. 256–259.

relationship between two categorical variables, in a two-dimensional contingency table.

Example 19.2 Relationship between color of the eyes and that of one hair Table 19.8 was a summary of color of the eyes and color of one hair of 5387 pupils from the Caithness County in northern Scotland. The objective was to study the correspondence relationship between the two. This was a 4×5 contingency table, used by Fisher in 1940 when he firstly introduced canonical analysis of contingency table data.

19.2.1 Computation Procedure

Suppose we have an $R \times C$ contingency table. Its rows and columns represent two categorical variables with R levels and C levels respectively. Let $\mathbf{X} = \{x_{ij}\}$ be the cell frequency in the table.

(1) Data transformation First we transform the original contingency table data into

$$z_{ij} = \frac{x_{ij} - \frac{x_{i.}x_{.j}}{x_{..}}}{\sqrt{\frac{x_{i.}x_{.j}}{x_{..}}}} \quad i = 1, 2, \dots, R : j = 1, 2, \dots, C. \quad (19.8)$$

Here $x_{i.}$ is the sum of i th row, and $x_{.j}$ is the sum of j th column, $x_{..}$ is the grand total. By using χ^2 -test, we know that x_{ij} is the observed frequency, $x_{i.}x_{.j}/x_{..}$ is the expected frequency under the null hypothesis that two categorical variables (row and column) are independent and the standardized residual

Table 19.9 Transformed values of z_{ij} from Table 19.8.

Color of the eyes	Color of the hair				
	Fair	Red	Medium	Dark	Black
Dark	-13.6444	-2.6129	-5.1964	18.5325	10.4736
Medium	-6.2167	-1.0496	7.736	-2.1505	-2.0624
Blue	9.4828	-0.0220	-2.5982	-5.5341	-3.2074
Light	12.6462	3.5083	-1.7101	-10.8920	-5.2038

Table 19.10 Color of the eyes (row effect) factor loadings.

Color of the eyes	First factor	Second factor	Raito of two factor loadings
Dark	-0.70274	0.13391	-5.24790
Medium	-0.03361	-0.24500	0.13718
Blue	0.40030	0.16541	2.42005
Light	0.44071	0.08846	4.98203

z_{ij} is essentially there,

$$z_{ij} = \frac{\text{Observed frequency} - \text{Expected frequency}}{\sqrt{\text{Expected frequency}}}$$

Data transform of the example yields the following Table 19.9.

(2) Calculation of two correlation matrices Utilizing transformed $R \times C$ data matrix $Z = \{z_{ij}\}$, we calculate the correlation coefficients between any two rows to obtain a correlation coefficient matrix **A**. Likewise, we calculate the correlation coefficients between any two columns to obtain a correlation coefficient matrix **B**. It can be shown that **A** and **B** have identical nonzero eigenvalues, but different eigenvectors. The example showed here has three nonzero eigenvalues, 0.1992, 0.03009 and 0.0008595. Their corresponding percentages of contribution are 86.56%, 13.07% and 0.37%.

(3) Perform a factor analysis based on **A**, we can get the row effect's factor loadings. The example here uses a two-factor model. The results are displayed in Table 19.10, where the last column is the ratio of two factor loadings.

Table 19.11 Color of the hair (column effect) factor loadings.

Color of the eyes	First factor	Second factor	Ratio of two factor loadings
Fair	0.54400	0.17384	3.12930
Red	0.23326	0.04828	4.83140
Medium	0.04202	-0.20830	-0.20173
Dark	-0.58871	0.10395	-5.66340
Black	-1.09439	0.28644	-3.82070

Table 19.12 The optimal correspondence of the contingency Table 19.8.

Color of the eyes	Color of the hair					
	Dark	Black	Medium	Fair	Red	Total
Dark	681	85	403	98	48	1315
Medium	412	26	909	343	84	1774
Blue	110	3	241	326	38	718
Light	188	4	584	688	116	1580
Total	1391	118	2137	1455	286	5387

(4) Perform a second factor analysis based on **B**, we can get the column effect's factor loadings. The current example again uses a two-factor model. The results are displayed in Table 19.11, where the last column is the ratio of two factor loadings as well.

Thus we have showed computations of correspondence analysis. Such analyses help reveal correspondence relationship between row and column effects.

19.2.2 The use of correspondence analysis

(1) Optimal correspondence Rearrange the order of each level in row and column effects according to the order of their factor loading ratio, from the highest to the lowest. In current example, the order of color of the eye remains unchanged, however, the effect level of color of the hair should be rearranged to Dark, Black, Medium, Fair, and Red. The optimal correspondence is therefore as in Table 19.12.

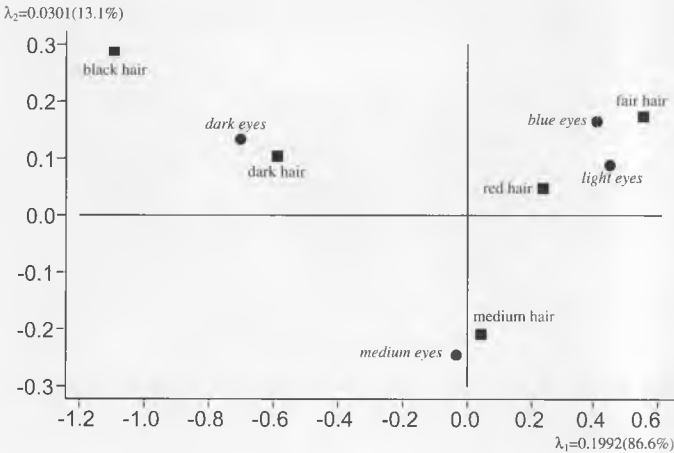


Fig. 19.1 Color of the eyes and color of the hair correspondence analysis of 5387 pupils.

Table 19.12 most effectively reveals the correlation between the color of the eyes and color of the hair, namely the color of the eyes from dark to light, corresponds to the color of the hair from dark to fair.

(2) Factor loading plot Analogous to factor loading plots in factor analysis, we plot the first factor against the second factor in a Cartesian system with factor loadings as the coordinates. In the plot, each level of row and column effects is identified to reveal correspondence relationship. Figure 19.1 is the factor loading plot for the above example, where the dots represent the color of the eyes (row effect), and the squares represent the color of the hair (column effect). It is easy to see that dark and black hair correspond to dark eyes, fair and red hairs correspond to blue and light eyes, and medium hair corresponds to medium eyes.

19.2.3 Application — Analysis of contraceptive methods used by various groups of people

Example 19.3 Wang Shaoxian surveyed 11,764 married women aged between 20 and 40 years in 1986 from nine cities in China, regarding their marriage, reproduction, and contraceptive uses. the summary data are showed in Table 19.13.

Make a transformation of (19.8), then calculate matrices **A** and **B**, to obtain four nonzero factors, with four eigenvalues 0.2686, 0.07685,

Table 19.13 Contraceptive methods frequency in nine cities (including urban and rural areas) in 1983.

		Contraceptive method					Total
Area		Intrauterine device (IUD)	Oral	Male condom	Sterilization	Other	
Urban	Beijing	153	33	165	40	40	431
	Jilin	346	10	15	76	10	457
	Chengdu	241	38	134	21	35	469
	Changsha	184	21	106	64	60	435
	Dalian	367	18	129	11	25	550
	Xi'an	703	55	130	69	83	1040
	Zhengzhou	248	12	113	60	30	463
	Chongqing	296	20	87	36	26	465
	Wuhan	476	79	113	82	91	841
Rural	Beijing	320	75	43	62	18	518
	Jilin	249	6	10	119	8	392
	Chengdu	278	38	22	141	36	515
	Changsha	73	4	13	323	10	423
	Dalian	209	43	66	100	7	425
	Xi'an	288	4	0	418	1	711
	Zhengzhou	141	6	1	294	1	443
	Chongqing	435	1	2	73	2	513
	Wuhan	364	164	4	277	16	825
Total		5371	627	1153	2266	499	9916

0.05063, and 0.01328. Their percentages of contribution are 65.62%, 18.77%, 12.37% and 3.24%, respectively. The first two factors contribute more than 80%. The factor loadings are showed in Tables 19.14 and 19.15.

Both row and column effects' on the first and second factors are plotted in Fig. 19.2. We could see, "sterilization" was far from urban groups, indicating urban residents' dislike of the method; "Male condom" and "Other" short term methods were far from rural groups, indicating the rural residents' preference. In Changsha, Zhengzhou, and Xi'an's rural area, "sterilization" was more popular than other cities' rural areas; other rural areas mostly used "IUD" and "oral contraceptive" methods. Furthermore, the plot characterizes different patterns of contraception use between urban and rural areas. The factor loadings form two data groups: the big ellipsis encompasses all rural areas, which contains "sterilization", "oral", and "IUD" three methods; the smaller ellipsis encompasses all urban areas, which contains "oral",

Table 19.14 Region (row effect) factor loading.

Region	Urban		Rural	
	First factor	Second factor	First factor	Second factor
Beijing	0.58482	0.66163	0.17327	-0.23431
Jilin	0.01999	-0.40534	-0.26735	-0.23758
Chengdu	0.56762	0.28204	-0.15591	-0.10461
Changsha	0.36027	0.40746	-1.19380	0.43462
Dalian	0.54516	0.01763	0.00283	0.07944
Xi'an	0.35924	-0.15586	-0.88458	0.05227
Zhengzhou	0.34615	0.23350	-1.03567	0.17254
Chongqing	0.38568	-0.01185	0.02638	-0.56771
Wuhan	0.32447	-0.00822	-0.35968	-0.15070

Table 19.15 Contraceptive method (Column effect) factor loading.

Contraceptive method	First factor	Second factor
IUD	0.15997	-0.21894
Oral	0.14711	-0.06410
Male condom	0.70451	0.57905
Sterilization	-0.89268	0.17539
Other	0.51924	0.30274

“IUD”, “male condom”, and “other” short term contraceptive methods. These two ellipses intersect with each other, indicating that “oral” and “IUD” were popular in most areas, whereas “sterilization” was used less in urban areas. In rural areas of Changsha, Zhengzhou, and Xi'an, “sterilization” was observed more; and in all considered rural areas, “male condom” and “other” short term methods were less popular.

19.2.4 Relationship between correspondence analysis and canonical correlation analysis

Canonical correlation analysis can also be applied to the data in Example 19.2. Table 19.8 is a contingency table, but not a multivariate data matrix, i.e., it is not a matrix with rows as observations and columns as variables. We first transform Table 19.8 into an n row and nd ($R + C$) column data matrix, as showed in Table 19.16. Here, the first R columns represent row effects (X variables), the next C columns represent column

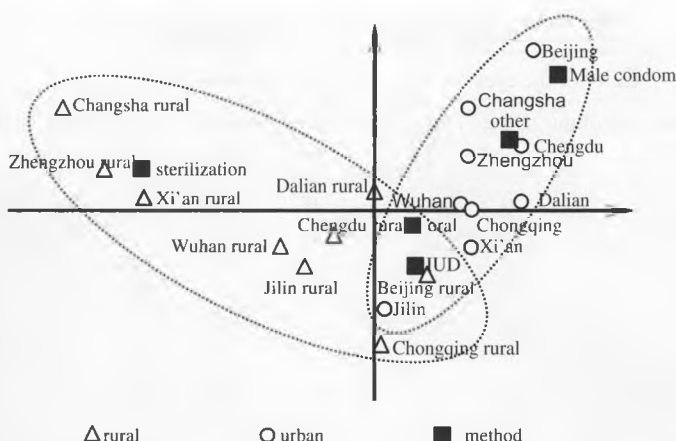


Fig. 19.2 Results of the correspondence analysis for Example 19.3.

effects (Y variables). The original ij th entry n_{ij} becomes, in this new data matrix, a "1" in the i th X variable, and a "1" in j th Y variable, the rest are all "0", with a frequency of n_{ij} . For instance, $n_{11} = 98$, in transformed data matrix becomes, $X_1 = 1$, $Y_1 = 1$, and the rest variables are all 0, with a frequency of 98. Likewise, $n_{23} = 909$ becomes $X_2 = 1$, $Y_3 = 1$, the rest are all 0, with a frequency of 909.

After transforming the data into a data matrix Table 19.16, we can perform canonical correlation analysis on X variables (row effects) and Y variables (column effects). Such an analysis leads to three canonical correlation coefficients 0.446368, 0.173455 and 0.029317. The squared correlation coefficients are 0.199245, 0.030087 and 0.000859, which are just the eigenvalues in the correspondence analysis. In fact, factors in correspondence analysis are the same as those in canonical correlation analysis. The main difference between the two is that canonical correlation analysis emphasizes on the canonical variables and canonical correlation coefficients, while correspondence analysis attempts to interpret the structure of contingency table.

19.3 Canonical Discriminant Analysis

Suppose there are k samples from $k \geq 2$ populations. Each sample contains observations from X_1, X_2, \dots, X_p , $p > k$, with samples size n_i ,

Table 19.16 Canonical correlation analysis of Example 19.2.

No.	Color of the eyes				Color of the hair					Frequency <i>f</i>
	Dark <i>X</i> ₁	Medium <i>X</i> ₂	Blue <i>X</i> ₃	Light <i>X</i> ₄	Fair <i>Y</i> ₁	Red <i>Y</i> ₂	Medium <i>Y</i> ₃	Dark <i>Y</i> ₄	Black <i>Y</i> ₅	
1	1	0	0	0	1	0	0	0	0	98
2	0	1	0	0	1	0	0	0	0	343
3	0	0	1	0	1	0	0	0	0	326
4	0	0	0	1	1	0	0	0	0	688
5	1	0	0	0	0	1	0	0	0	48
6	0	1	0	0	0	1	0	0	0	84
7	0	0	1	0	0	1	0	0	0	38
8	0	0	0	1	0	1	0	0	0	116
9	1	0	0	0	0	0	1	0	0	403
10	0	1	0	0	0	0	1	0	0	909
11	0	0	1	0	0	0	1	0	0	241
12	0	0	0	1	0	0	1	0	0	584
13	1	0	0	0	0	0	0	1	0	681
14	0	1	0	0	0	0	0	1	0	412
15	0	0	1	0	0	0	0	1	0	110
16	0	0	0	1	0	0	0	1	0	188
17	1	0	0	0	0	0	0	0	1	85
18	0	1	0	0	0	0	0	0	1	26
19	0	0	1	0	0	0	0	0	1	3
20	0	0	0	1	0	0	0	0	1	4

Table 19.17 Definition of indicator variables.

Group	<i>Y</i> ₁	<i>Y</i> ₂	...	<i>Y</i> _{<i>k</i>-1}
1	1	0	...	0
2	0	1	...	0
⋮	⋮	⋮	⋮	⋮
<i>k</i> - 1	0	0	...	1
<i>k</i>	0	0	...	0

$i = 1, 2, \dots, k$. We wish to use these samples as our training sample, to derive a discriminant rule. This is a generalization of two-population discrimination problem.

Define $k - 1$ indicator variables Y_1, Y_2, \dots, Y_{k-1} , which take values 0 or 1 according to Table 19.17.

For each individual, in addition to X_1, X_2, \dots, X_p , there are "observations" Y_1, Y_2, \dots, Y_{k-1} . Therefore, we can derive the first pair of standardized canonical variables

$$\begin{aligned} W_1 &= a_{11}Y_1 + a_{12}Y_2 + \dots + a_{1,k-1}Y_{k-1}, \\ V_1 &= b_{11}X_1 + b_{12}X_2 + \dots + b_{1p}X_p. \end{aligned}$$

V_1 is called the first standardized canonical discriminant function, because it has the highest correlation with the indicator variables.

Similarly, we can obtain W_2 and V_2 . V_2 is called the second canonical discriminant function, which is uncorrelated to V_1 .

Proceeding accordingly, we can obtain V_3, V_4, \dots, V_{k-1} , a total of $k-1$ canonical discriminant functions. When $k=2$, there is only one discriminant function, which is commonly referred to as Fisher's discriminant function. It can be showed that the canonical discriminant functions are multi-population Fisher's discriminant functions.

Similar to two-population problem, we have only obtained discriminant functions without a discriminant rule. In general, we derive discriminant rules based on the values of V_1, V_2, \dots, V_{k-1} . The simplest approach is to use the following minimal distance discrimination procedure.

- (1) Calculate the values of V_1, V_2, \dots, V_{k-1} for each individual;
- (2) Within each sample, calculate the mean value of V_1, V_2, \dots, V_{k-1} , as its gravity center;
- (3) Calculate new observation's (of unknown group) discriminant function values, denoted by $V_1^*, V_2^*, \dots, V_{k-1}^*$.
- (4) Calculate the Euclidean distances of discriminant function values of (3) from each of the k gravity centers in (2). The individual is allocated to the group, of which the gravity center has the shortest distance to it.

The discrimination may be improved, if we use a weighted Euclidean distance instead of the Euclidean distance in (4). The weight w_i can be the reciprocal of the variance of V_i 's within a sample. That is,

$$w_i = [\text{var}(V_i)]^{-1}, \quad d_j^2 = \sum_{i=1}^{k-1} w_i (V_i^* - \bar{V}_i)^2.$$

Here d_j^2 is the (squared) distance between $V_1^*, V_2^*, \dots, V_{k-1}^*$ and the i th gravity center $(\bar{V}_{1j}, \bar{V}_{2j}, \dots, \bar{V}_{k-1,j})$.

19.4 Computerized Experiments

Experiment 19.1 Canonical correlation analysis Using the 1985 physical survey data of male students from 28 Chinese cities (ages 19–22) to perform a canonical correlation analysis.

Program 19.1 Canonical correlation analysis.

Line	Program	Line	Program
01	DATA A;	07	77.7 113.3 72.1 52.8 4238
02	INPUT X1-X6 Y1-Y5;	08	;
03	CARDS;	09	PROC CORR OUT=CORREL;
04	173.28 93.62 60.10 86.72 38.97 27.51	10	PROC CANCORR DATA= CORREL ALL;
05	75.3 117.4 74.6 6108 4508	11	VAR X1-X6;
06	168.99 91.52 51.11 86.23 38.30 27.14	12	WITH Y1-Y5;
		13	RUN;

In Program 19.1, lines 01–08 are data entry. Line 09 does correlation analysis among the variables, and stores correlation matrix into dataset CORREL. Lines 10–13 do canonical correlation analysis, analyzing relationship between variables X1 through X6 and Y1 through Y5.

Experiment 19.2 Correspondence analysis Perform a correspondence analysis on the example of pupils' color of the eyes and color of the hair relationship. Data are in Table 19.8.

Program 19.2 Correspondence analysis.

Line	Program	Line	Program
01	DATA A;	11	VAR Fair Red Medium Dark Black;
02	INPUT eye \$ Fair Red Medium Dark Black;	12	ID eye;
03	CARDS;	13	RUN;
04	Lighteye 688 116 584 188 4	14	PROC PLOT DATA=result;
05	Blueeye 326 38 241 110 3	15	WHERE eye NE "";
06	Mediumeye 343 84 909 412 26	16	PLOT DIM2*DIM1="*" \$ eye /BOX VAXIS=-.3 TO .3 BY .1 HAXIS=-1.2 TO 1 BY .2;
07	Darkeye 98 48 403 681 85	17	RUN;
08	;		
09	RUN;		
10	PROC CORRESP OUT=result;		

Lines 01–09 in Program 19.2 are again for data entry. Lines 10–13 perform correspondence analysis. Lines 14–17 make plots.

19.5 Practice and Experiments

1. The data on 1985 male students (19–22 years of age) from 28 Chinese cities are summarized in Table 19.18 (the city numbers are the same as those used previously). Using morphology indices and function indices described in Chap. 17 to perform the following tasks:

Table 19.18 The physical ability data of male students (19–22) from 28 Chinese cities.

City	50 meters sprint (sec.) Z_1	Stationary jump (cm) Z_2	Pull-up (times) Z_3	Standing forward bending (cm) Z_4	100 meters sprint (sec.) Z_5
1	7.48	225.3	5.9	8.01	248.21
2	7.63	218.0	3.9	10.25	248.51
3	7.49	226.1	7.7	13.56	235.08
4	7.83	216.2	6.2	9.98	257.53
5	7.46	232.7	7.4	14.14	228.07
6	7.42	232.7	7.6	12.91	226.81
7	7.58	220.4	7.9	12.61	236.56
8	7.32	229.8	7.5	10.51	237.78
9	7.30	229.5	8.3	11.78	229.88
10	7.62	223.4	7.5	12.56	238.65
11	7.59	216.5	6.7	13.23	247.43
12	7.73	219.0	6.4	9.83	246.49
13	7.43	225.7	7.4	12.61	242.96
14	7.57	224.6	7.7	13.62	245.92
15	7.39	227.8	7.6	7.14	246.10
16	7.23	231.6	9.0	11.45	228.73
17	7.34	227.8	8.3	12.05	233.06
18	7.64	218.3	7.1	12.75	237.52
19	7.46	222.7	6.7	9.34	242.39
20	7.36	239.1	10.3	11.70	225.13
21	7.53	225.0	9.4	13.53	237.37
22	7.41	233.0	8.1	11.09	234.12
23	7.57	227.1	9.0	13.22	231.09
24	7.15	228.1	9.5	9.77	223.09
25	7.35	237.5	10.4	13.20	224.34
26	7.43	225.2	7.7	10.06	236.67
27	7.69	218.9	7.7	12.15	243.67
28	7.37	224.1	7.4	11.27	245.32

- (1) Canonical correlation analysis of X_1, X_2, \dots, X_6 and Z_1, Z_2, \dots, Z_5 ;
 - (2) Canonical correlation analysis of Y_1, Y_2, \dots, Y_5 and Z_1, Z_2, \dots, Z_5 ;
 - (3) Canonical correlation analysis of $X_1, X_2, \dots, X_6, Y_1, Y_2, \dots, Y_5$ and Z_1, Z_2, \dots, Z_5 .
 - (4) Discuss the difference between performing (1) and (2) separately versus a combined analysis of (3). In your opinion, which is preferred and why?
2. Suppose there are two sets of variables X_1, X_2, \dots, X_p and Y_1, Y_2, \dots, Y_q .
- (1) When $p = q = 1$, can you use canonical correlation analysis to obtain the simple correlation coefficients and simple linear regression?
 - (2) When $q = 1$, can you use canonical correlation analysis to obtain the multiple correlation coefficients and multiple linear regression?
3. Use canonical correlation analysis programs to solve Problem 2 in Exercises of Chap. 18. Then compare the results.
4. Use principal component analysis to find the principal components of X and Y variables in Problem 1 above. Then compute correlation coefficients between the principal components of X and Y variables. Compare and discuss the results with that of Problem 1.
5. Combine morphology data in Table 18.1 and functional data in Table 18.8 to perform a factor analysis. Use the factor model obtained to identify highly correlated X and Y variables. Compare with Table 19.5, and discuss.
6. Table 19.19 is a summary of the analgesic effect of pain relievers. Perform a correspondence analysis on the data.

Table 19.19 Data on the analgesic effect of pain relievers.

Drug	Analgesic effect				
	Poor	Fair	Good	Very good	Excellent
A	5	1	10	8	6
B	5	3	3	8	12
C	10	6	12	3	0
D	7	12	8	1	1

Source: Michael J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press. 1984, 263.

(1st edn. Fei Lin, Ying Lu, Jie Yan; 2nd edn. Yuanto Hao, Nanqiao Cai, Jiqian Fang)

Chapter 20

Survival Analysis

To determine the prognosis circumstance of diseases, we must not only look at the final outcome being stand or fall, but also observe the time length for experiencing this kind of final outcome. For example the prediction that in Table 20.1, patients Nos. 2, 4 and 5 all die, but the lengths of life time are different, and namely the prognosis of these three patients are different. Generally such data are collected by follow-up, and the begin time is the diagnosis date, treatment date etc. The clearest definition of the final outcome is death, but in addition, is the relapse, the deformity, recover from illness etc. The data of follow-up is often incomplete because of lost to visit, and hence a kind of specialized methods is needed for statistical analysis, which are usually called survival analysis or analysis of life data. Survival analysis is also suitable to the research to track factors of diseases (occurring disease as positive), or to track clinical curative effect (recover from illness or remarkable effect as positive), or to the animal examination (disease or death as positive) etc.

It is worthy to note that using percentage index, such as cure rate, effective rate, fatality rate or disability rate to evaluate clinical effect is not very ideal. For example, a doctor treats 50 gastric ulcer patients by using the herb and the routine medicine respectively, and the recovery rates from illness are 90% in both cases. It seems that the curative effect for these two kinds of drugs is the same. But the assistant of that doctor discovers that the gastric ulcer disappeared in 20 days on average for the group with herb, and is 30 days on average in another group, so it seems that the herb is better than the routine medicine. This shows that the curative effect described only by percentage is unilateral and crude. The time factor must also be considered when analyzing curative effect. The survival analysis is the method to evaluate the curative effect completely and accurately.

Table 20.1 Follow-up data of 5 cases with liver cancer.

No.	Name	Covariates		Observed records			
		Gender (M = 1)	Group	Begin date	End date	Outcome (dead = 1)	Survival days
1	MSL	1	0	98-07-12	98-11-29	0	140
2	LSL	0	1	98-07-01	98-12-08	1	160
3	ZXJ	1	1	98-07-14	98-12-31	0	170
4	WYQ	0	0	98-08-22	98-11-29	1	99
5	WJS	1	1	98-10-20	98-11-25	1	36

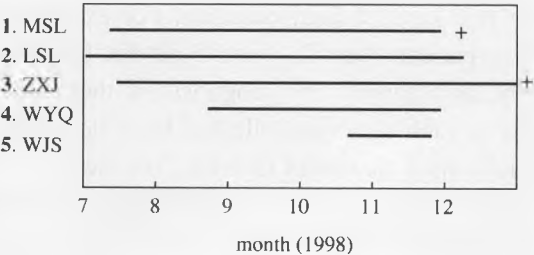


Fig. 20.1 Sketch map of survival times in original records (the “+” means still living, the same as follows).

20.1 The Basic Concept of Survival Analysis

20.1.1 The record of follow-up data (data construction)

Example 20.1 The follow-up data of five cases are showed in Table 20.1, where the records include the beginning date to observe, terminal date, final outcome and covariates (study factor and confounder). The characteristics of this kind of data are:

- (1) Two dependent variables, namely the survival time (days) and the final outcome (death or not);
- (2) The survival time may be observed incompletely, therefore we do not know how long the patient will actually live, such as patients Nos. 1 and 3 in Table 20.1. In order to reflect the effects of prognosis, the correct method must incorporate the final outcome with survival time by survival analysis.

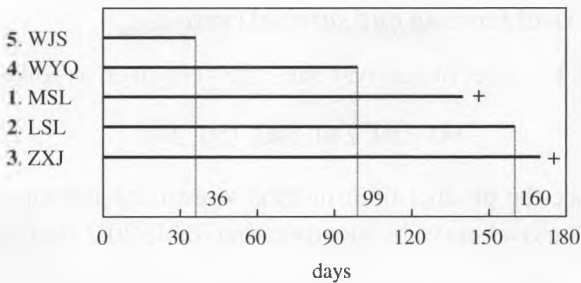


Fig. 20.2 Sketch map of survival times being sorted.

20.1.2 Survival time complete data, censored data

The survival time means the length of survival time observed, such as 140, 160, 170, 99, 36 days in Table 20.1. There are two kinds of survival time as follows:

- (1) The complete data: The time from beginning time to the time of death, namely survival time of death cases, such as 160, 99 and 36 days in Table 20.1.
- (2) The censored data: Because of failure to follow up, changing the plan of treatment or ending the research etc., some patients cannot be followed to their death, which is called with censoring or being censored. The time length from beginning point to censoring point, namely the survival time of the survivor, is called censored data, such as 140, 170 days in Table 20.1, customarily recorded as 140^+ , 170^+ .

The complete data provide the accurate times of survival, and they are the basis of analysis. The censored data also provide some information that the patient still survive at the censored time.

20.2 The Product-Limit Method for One Group of Survival Data

This method directly uses the multiplication of survival probabilities corresponding to the complete death times to estimate the survival function so that it is called the product-limit method. It was proposed by Kaplan–Meier (1958). This is a kind of non-parameter method in nature.

20.2.1 Survival function and survival curve

Example 20.2 A set of survival data (days) is given as follows:

$$90, 150, 210, 540, 150, 270^+$$

Now introduce the product-limit method to estimate the survival function and plot the survival curve by incorporating Table 20.2 step by step:

- (1) Sorting the observed time points (t). List the completed times in increasing order (Column 2); not including the censored data (such as “270+”); if there is a tie, the same value appears just once (such as “150”).
- (2) Record the number of deaths at each time point (Column 3).
- (3) Record the number of censored ones at the observed time point right before the censored time (Column 4). For example, for the censored time 270^+ , we count it into the point of 210, for $210 < 270 < 540$.
- (4) Calculate the number of cases right before the time point t (Column 5). In this case, the number of cases right before the first time point 90 is $n_{01} = 6$; for other time points, one may get the number of cases by subtraction, such as

$$n_{02} = n_{01} - d_1 - c_1 = 6 - 1 - 0 = 5$$

$$n_{03} = n_{02} - d_2 - c_2 = 5 - 2 - 0 = 3.$$

Alternatively, the number of cases right before the last time point is $n_{04} = 1$; for other time points, one may accumulate backward to get the number of cases, such as

$$n_{03} = n_{04} + d_3 + c_3 = 1 + 1 + 1 = 3,$$

$$n_{02} = n_{03} + d_2 + c_2 = 3 + 2 + 0 = 5.$$

- (5) Calculate the probability of death and probability of survive at each time point (Columns 6 and 7).

$$q_k = \frac{d_k}{n_{0k}}, \quad p_k = 1 - q_k. \quad (20.1)$$

- (6) Calculate the survival function at each time point (Column 8).

$$S(t_k) = P(T \geq t_k) = p_1 p_2 \cdots p_k. \quad (20.2)$$

Table 20.2 Calculation of survival function by Kaplan–Meier method.

No. (1)	Time to death (2)	Number of deaths at t (3)	Number of censored at t (4)	Number of cases right before t (5)	Probability of death at t (6)	Probability of survive at t (7)	Survival function (8)	Standard error of $S(t)$ (9)
0	0	0	0	6	0	1	1	
1	90	1	0	6	1/6	5/6	$(5/6) = 0.833$	0.152
2	150	2	0	5	2/5	3/5	$(5/6)(3/5) = 0.500$	0.204
3	210	1	1	3	1/3	2/3	$(5/6)(3/5)(2/3) = 0.333$	0.193
4	540	1	0	1	1/1	0/1	$(5/6)(3/5)(2/3)(0/1) = 0$	0

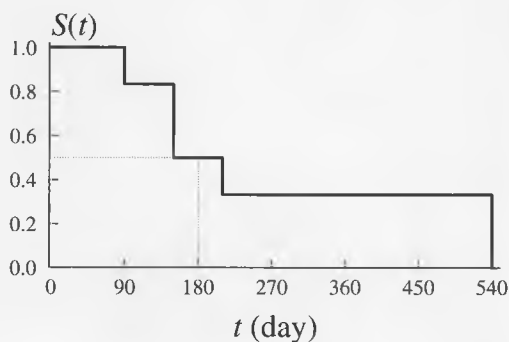


Fig. 20.3 Survival curve made by K-M method and its median ($Md = 180$ days).

- (7) Calculate the standard error of the survival function (Column 9). The formula was proposed by Greenwood (1926)

$$SE(S(t_k)) = S(t_k) \sqrt{\sum_{j=1}^k \frac{q_j}{p_j n_j}}. \quad (20.3)$$

- (8) Plot the survival curve according to columns 2 and 8 (Fig. 20.3).

For product-limit method, the survival function is estimated at each time point of death, and assumed as a right continuous step function. Therefore its graph looks like a stair. The value of survival function at time point t is at down a step, such as $S(90) = 0.8333$ in Fig. 20.3, instead of 1. If there is no case censored at the last time point, the curve will end with the x -axis (disadvantage). When both the sample size and the number of death time points are large, the stair-form of the survival curve will not be obvious.

Based on survival curve, we can estimate the median-lifetime and quartile range (Fig. 20.3).

20.2.2 The basic requirements by K-M method

- (1) A random sample with big enough sample size is available;
- (2) The number of deaths should not be too small (say, ≥ 30);
- (3) The proportion of the censored data should not be too large.

Table 20.3 The relapse times (month) for children with rhabdosarcoma after treatments*.

Control	2	3	9	10	10	12 ⁺	15	15 ⁺	16	18 ⁺	24 ⁺	30	36 ⁺	40 ⁺	45 ⁺
Treatment	9	12 ⁺	16	19	19 ⁺	20 ⁺	20 ⁺	24 ⁺	24 ⁺	30 ⁺	31 ⁺	34 ⁺	42 ⁺	44 ⁺	53 ⁺
	59 ⁺	62 ⁺													

*“+” means no relapse yet.

20.3 The Log-Rank Test and Breslow Test for Comparing Two Survival Data Sets

In this section, the tests for comparing two groups of survival data will be emphasized. The null hypothesis H_0 is that the two survival functions are exactly the same although the whole process of test generally does not estimate the survival function.

Example 20.3 There are two sets of relapsed times (month) for children with rhabdosarcoma after treatments as shown in Table 20.3. The control group is “extirpate + actinotherapy”, and treatment group is “extirpate + actinotherapy + chemotherapy”. Whether the chemotherapy on the basis of “extirpate + actinotherapy” can increase the remission rate? (similar to survival rate).

Now let us introduce two tests for this kind of problems incorporating Example 20.2. The two remission curves in Fig. 20.4 are estimated with the above introduced product-limit method. The difference between them needs to be tested with the following hypotheses:

H_0 : The two remission curves are the same in population

H_1 : The two remission curves are not the same in population

First, uncensored relapse time points are sorted by mixing the data of two groups, shown in columns 1–5 in Table 20.4. The table consists of many fourfold tables, each of which corresponding to a time point of relapse. Promising for the j th fourfold table, the total numbers of exposed, relapse and non-relapse are denoted with N_j , D_j and S_j respectively, and for the control (or the treatment) group, the number of exposed and relapse are denoted with n_j and d_j respectively. When H_0 is true, the a_j in the fourfold

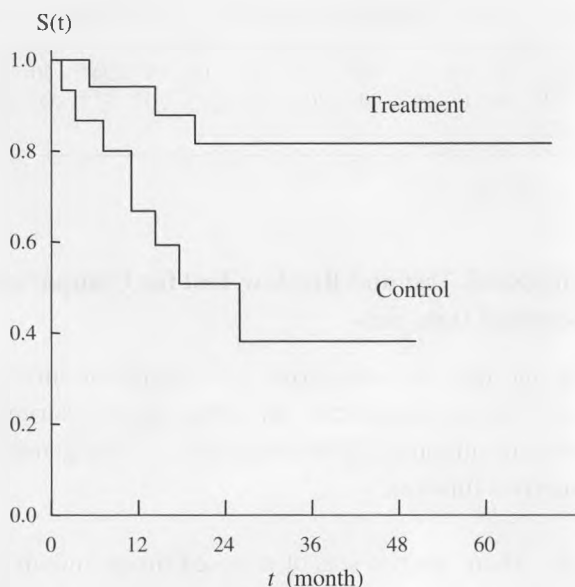


Fig. 20.4 Comparison between the two remission curves for the treatment group and the control group.

table is a random variable (the rest are determined by a_j , N_j and D_j), of which its expected number and variance are

$$e_j = n_j \frac{D_j}{N_j}, \quad v_j^2 = \frac{D_j S_j n_j (N_j - n_j)}{N_j^2 (N_j - 1)}.$$

When there are enough number of such kind of fourfold tables, we have the sum of the expected frequencies $\sum e_j$, the sum of the observed frequencies $\sum a_j$ and the corresponding variance $\sum v_j^2$.

20.3.1 The log-rank test

The statistics of log-rank test is

$$\chi_L^2 = \frac{(\sum a_j - \sum e_j)^2}{\sum v_j^2}, \quad (20.4)$$

when H_0 is true, this statistic follows a χ^2 distribution with one degree of freedom.

Table 20.4 Calculation for log-rank test and Breslow test.

j	Time to relapse (month)	Contr.	Treat.	Sub total	$e = a - \frac{nD}{N}$	$v^2 = \frac{DS_{n(N-n)}}{N^2(N-1)}$	Na	$Ne = nD$	N^2v^2
1	2	1(a, e) 14	0 7	1 (D) 31 (S)	0.469	0.249	32	15	254.976
		15 (n)	17 (N - n)	32 (N)					
2	3	1 13	0 17	1 30	0.452	0.248	31	14	238.328
		14	17	31					
3	9	1 12	1 16	2 28	0.867	0.474	30	26	426.600
		13	17	30					
4	10	2 10	0 16	2 26	0.857	0.472	56	24	370.048
		12	16	28					
5	15	1 8	0 15	1 23	0.375	0.234	24	9	134.784
		9	15	24					

(Continued)

Table 20.4 (Continued)

Time to relapse									
j	(month)	Contr.	Treat.	Sub total	$e = a - \frac{nD}{N}$	$v^2 = \frac{DS_{n(N-n)}}{N^2(N-1)}$	Na	$Ne = nD$	N^2v^2
6	16	1	1	2	0.636	0.413	22	14	199.892
		6	14	20					
		7	15	22					
7	19	0	1	1	0.263	0.194	0	5	70.034
		5	13	18					
		5	14	19					
8	30	1	0	1	0.333	0.222	12	4	31.968
		3	8	11					
		4	8	12					
Total		8	3		4.252	2.506	207	111	1726.630
		$\sum a_j$			$\sum e_j$	$\sum v_j^2$	$\sum N_j a_j$	$\sum N_j e_j$	$\sum N_j^2 v_j^2$

Example 20.4 Perform a log-rank test for the data of Example 20.3.

Solution Referring to Table 20.4,

$$\sum a_j = 8, \quad \sum e_j = 4.252, \quad \sum v_j^2 = 2.506.$$

Substitute into (20.4)

$$\chi_L^2 = \frac{(8 - 4.252)^2}{2.506} = 5.605.$$

According to the χ^2 distribution with one degree of freedom, $P = 0.0179$. H_0 is rejected so that it can be concluded that the remission rate might be increased by chemotherapy addition to "extirpate + actinotherapy".

20.3.2 Breslow test

For $j = 1, 2, \dots$, if a_j is weighted by exposed number N_j at the j th fourfold table, then Breslow's statistic is

$$\chi_B^2 = \frac{(\sum N_j a_j - \sum N_j e_j)^2}{\sum N_j^2 v_j^2}. \quad (20.5)$$

When H_0 is true, this statistic follows a χ^2 distribution with one degree of freedom.

Example 20.5 Perform the Breslow test for the data of Example 20.4.

Solution Referring to Table 20.4,

$$\sum N_j a_j = 207, \quad \sum N_j e_j = 111, \quad \sum N_j^2 v_j^2 = 1726.630.$$

Substitute into (20.5)

$$\chi_B^2 = \frac{(207 - 111)^2}{1726.630} = 5.338.$$

According to the χ^2 distribution with one degree of freedom, $P = 0.0209$. H_0 is also rejected so that it can be concluded that the remission rate might be increased by chemotherapy addition to "extirpate + actinotherapy".

20.4 The Cox Regression

20.4.1 A brief introduction

The Cox regression is one of the most important methods for the analysis of life data. The primary applications are prognostic analysis of tumor and other chronic diseases, and cohort study to explore causes of diseases.

20.4.1.1 The Cox model

Suppose there are p covariates X_1, X_2, \dots, X_p , the hazard function of death at time t is assumed as

$$h(t) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p). \tag{20.6}$$

Here $h_0(t)$ is the baseline hazard function depending on time t only. This model is proposed by Cox (1972), and also called proportional hazards model, which means that the hazard is proportional to $h_0(t)$, and the ratio between $h(t)$ and $h_0(t)$ is $e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}$ or $\exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)$, depending on the individual's values of X_1, X_2, \dots, X_p .

To illustrate the model in detail, the data of four cases of liver cancer are listed on the left of Fig. 20.5, and the value of hazard function is demonstrated on the right of the figure. One can see that the hazard function varies with the individual's values of covariates given on the left-hand side, and also varies with time for the same individual. The Cox regression model is subject to a semi-parametric model because only the part of

Name of patient	Group X_1	Gender (M=1) X_2	Days t	Issue (D=1) d	Hazard function (different from one to another) $h(t)=h_0(t)e^{\beta_1 X_1 + \beta_2 X_2}$	Value of hazard (vary with t)			
						36 days	99 days	140 days	180 days
A	1	1	36	1	$h_0(t)e^{\beta_1 + \beta_2}$	$h_0(36)e^{\beta_1 + \beta_2}$			
B	0	0	99	1	$h_0(t)$	$h_0(36)$	$h_0(99)$		
C	0	1	140	0	$h_0(t)e^{\beta_2}$	$h_0(36)e^{\beta_2}$	$h_0(99)e^{\beta_2}$	+	
D	1	0	180	1	$h_0(t)e^{\beta_1}$	$h_0(36)e^{\beta_1}$	$h_0(99)e^{\beta_1}$		$h_0(180)e^{\beta_1}$

Fig. 20.5 The sketch map about the model of Cox regression (four cases of liver cancer).

$\exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)$ is specified as an exponential function with parameters $\beta_1, \beta_2, \dots, \beta_p$, but $h_0(t)$ is not specified.

20.4.1.2 *Estimation and test for regression coefficients*

Because the form of $h_0(t)$ is not specified, the traditional maximum likelihood estimation does not work for Cox regression. Cox (1972, 1975) proposed a partial likelihood function which is the product of conditional mortality probabilities at the points that deaths happened; he suggested to estimate the parameters by maximizing this partial likelihood function; and to perform the tests for the parameters by score test, Wald test and likelihood ratio test. The optimal properties of the estimates and the rationale of the tests have been studied by many statisticians since then. Here the theory of parameter estimation is skipped and the algorithm can be performed by statistical software, which will be introduced later.

20.4.2 *Application of Cox regression*

We would introduce the main steps of the application of Cox regression incorporating with an example as follows:

Example 20.6 The curative effect of variola powder to chorion carcinoma has been studied. 16 nude mice, who were successfully inoculated with chorion carcinoma at their body surface, were randomly divided into four groups to accept four treatments respectively (none, variola powder, drug A and drug B). The data are showed in Table 20.5. Work out the statistical analysis by Cox regression.

20.4.2.1 *Data structure*

As those in Table 20.5, the data required by Cox regression include at least four parts: begin date, terminal date, outcome (dead or not), covariates. The first three are necessary.

20.4.2.2 *Preliminary screening of the covariates*

Although the number of covariates is unlimited, if it is too much or the quality of data on covariates is not good, a preliminary screening of covariates is necessary. In general, eliminate the covariates with more missing

Table 20.5 The curative effect of variola powder to chorion carcinoma.

No.	Covariate (factor)						Observations			
	Days with tumor td	Size of tumor v0	Variola powder tr1	Drug A tr2	Drug B tr3	Vitamin C vitC	Begin date date0	End date date1	Death d	Survival time (days) day
1	19	25	0	0	0	1	89-05-20	89-05-28	1	8
2	17	16	0	0	0	1	89-05-20	89-05-29	1	9
3	19	37	0	0	0	1	89-05-20	89-05-28	1	8
4	16	19	0	0	0	1	89-05-20	89-05-28	1	8
5	14	25	1	0	0	1	89-05-20	89-06-07	0	18
6	13	18	1	0	0	1	89-05-20	89-06-06	1	17
7	16	25	1	0	0	1	89-05-20	89-06-03	1	14
8	9	10	1	0	0	1	89-05-20	89-06-04	1	15
9	9	22	0	1	0	1	89-05-20	89-06-04	1	15
10	10	25	0	1	0	1	89-05-20	89-05-31	1	11
11	14	25	0	1	0	1	89-05-20	89-06-02	1	13
12	12	37	0	1	0	1	89-05-20	89-06-01	1	12
13	17	37	0	0	1	1	89-05-20	89-05-29	1	9
14	14	29	0	0	1	1	89-05-20	89-06-01	1	12
15	13	13	0	0	1	0	89-05-20	89-06-01	1	12
16	17	31	0	0	1	1	89-05-20	89-05-30	1	10

values or with too small variability (such as *vitC* in Table 20.7), because these covariates provide less information and might cause big trouble in computation.

Then each covariate is analyzed independently by Cox regression model with single covariate only. Eliminate those covariates that the corresponding *P* values are too big, because their coefficients in the model might be zero.

20.4.2.3 Cox regression model with multiple covariates

Take over all the statistically significant covariates in the above step into the Cox model to perform a Cox regression with multiple covariates.

Similar to the procedures for selection of variables in multiple linear regression, the forward or backward methods are in common use. Some times, the preferential method can be used as well, that is, the important

covariates could be selected into or kept staying in the model according to the professional knowledge.

Solution (of Example 20.6) Table 20.6 gives the results of the analysis for the data in Table 20.5 on the basis of Cox model with single covariate. The first column lists the covariates; columns 2 to 6 are the statistical descriptions for the covariates, including number of cases, maximum, minimum, mean and standard deviation; columns 7 to 9 are the tests for the Cox model with single covariate, whether each of the coefficients is zero or not; columns 7 and 8 list the values of test statistic χ^2 and the degrees of freedom; column 9 lists the P values, of which the larger values indicate the coefficients in the model might be zero or the corresponding covariates are not statistically significant. It shows that the covariates *td* and *tr1* are statistically significant while *tr2*, *tr3* and *v0* are not.

Table 20.7 gives the results on the basis of Cox model with two covariates *td* and *tr1*. Columns 2 and 3 are the estimated coefficients and their standard errors; columns 4 to 6 are the tests for the Cox model with two covariates, whether each of the coefficients is zero or not. It shows that both *td* and *tr1* are significantly different from zero. Therefore, we conclude that the curative effect of variola powder to chorion carcinoma is to decrease the hazard of death and increase the survival time, but the days with tumor might increase the hazard of death.

Table 20.6 The results by Cox regression with single covariate.

Covar. (1)	Statistical description					Test for model with single covariate		
	Case (2)	Max. (3)	Min. (4)	Mean (5)	SD (6)	χ^2 (7)	<i>df</i> (8)	<i>P</i> (9)
d	16	0.00	1.00	0.9375	0.2500			
day	16	8.00	18.00	11.9375	3.2139			
td	16	9.00	19.00	14.3125	3.2191	4.7033	1	0.0301
tr1	16	0.00	1.00	0.2500	0.4472	6.6318	1	0.0100
tr2	16	0.00	1.00	0.2500	0.4472	0.0091	1	0.9238
tr3	16	0.00	1.00	0.2500	0.4472	1.1985	1	0.2736
v0	16	10.00	37.00	24.6250	8.2694	1.4238	1	0.2328

Table 20.7 The results by Cox regression with multiple covariates.

Covar. (1)	Parameter		Test for model with multiple covariates			Risk ratio (7)	Standard parameter	
	Coefficient (2)	SE (3)	χ^2 (4)	df (5)	P (6)		Coefficient (8)	SE (9)
td	0.4201	0.1630	6.6467	1	0.0099	1.5221	1.3524	0.5246
tr1	-2.9399	1.0714	7.5297	1	0.0061	0.0529	-1.3148	0.4790

20.4.2.4 Risk ratio

Column 7 in Table 20.7 gives the risk ratio of each covariate. It is defined as how risk it will be while the covariate increases by one unit.

Let us take the risk ratio of X_1 as an example:

For $X_1 = x_1, X_2 = x_2, \dots, X_p = x_p$, the hazard function is

$$h_1(t) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p).$$

For $X_1 = x_1 + 1, X_2 = x_2, \dots, X_p = x_p$, the hazard function is

$$h_2(t) = h_0(t) \exp[\beta_1(x_1 + 1) + \beta_2 x_2 + \dots + \beta_p x_p].$$

Comparing the two hazard functions, one has

$$\frac{h_2(t)}{h_1(t)} = \frac{h_0(t) \exp[\beta_1(x_1 + 1) + \beta_2 x_2 + \dots + \beta_p x_p]}{h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)} = \exp(\beta_1).$$

Therefore, when the value of X_1 changes from x_1 to $x_1 + 1$, the hazard becomes $\exp(\beta_1)$ times of that before changing.

In general, for the i th covariate, we denote such a risk ratio with

$$RR_i = \exp(\beta_i). \quad (20.7)$$

If X_i is a binary variable, obviously, the risk ratio is exactly the same as the relative risk comparing the condition of $X_i = 1$ to that of $X_i = 0$.

Column 7 in Table 20.7 are calculated by (20.7),

$$RR_2 = \exp(\beta_2) \approx \exp(-2.9399) = 0.0529,$$

$$RR_1 = \exp(\beta_1) \approx \exp(0.4201) = 1.5221.$$

20.4.2.5 Standard coefficient

Columns 8 and 9 in Table 20.7 are the standard regression coefficient b'_i and its standard error $SE(b'_i)$ for i th covariate. As for the relationship between the regression coefficient b_i and b'_i , we have

$$b'_i = b_i S(X_i), \quad SE(b'_i) = SE(b_i) S(X_i). \quad (20.8)$$

Here $S(X_i)$ refers to the standard deviation of the covariate X_i .

Same as that in multiple linear regression, the standard regression coefficient b'_i can be used to compare the contributions of different covariates free of the problem of units. From column 8 of Table 20.7, one can see that the contribution of the two covariates *td* and *tr1* are about the same, while is not obvious from column 2.

20.4.3 The estimation of survival function

After the regression coefficients being estimated, one can go further to estimate the survival function by the method proposed by Breslow.

First of all, let us estimate the baseline survival function at i th death time t_i , denoted with $S_0(t_i)$

$$S_0(t_i) = \prod_{j=1}^i \left[1 - \frac{d_j}{\sum_{k|t_k \geq t_j} \exp(b_1 x_1^{(k)} + b_2 x_2^{(k)} + \dots + b_p x_p^{(k)})} \right]. \quad (20.9)$$

Here d_j is the number of deaths at j th death time t_j ; $(x_1^{(k)}, x_2^{(k)}, \dots, x_p^{(k)})$ refers to the k th individual's values of covariates; $\sum_{k|t_k \geq t_j}$ refers to a summation over all the patients who survived at the moment right before t_j ; $\prod_{j=1}^i$ refers to a product over t_1, t_2, \dots, t_j .

Comparing (20.9) to (20.1) and (20.2), one will find that, the only difference is that instead of simply the number of patients who survived at the moment right before t_j , here in the numerator is the sum of the "weights" of $\exp(\bullet)$.

Obviously, under the assumption of (20.6), the survival function of a specific individual with $X_1 = x_1^*, X_2 = x_2^*, \dots, X_p = x_p^*$ is equal to

$$S(t) = [S_0(t)]^{\exp(\beta_1 x_1^* + \beta_2 x_2^* + \dots + \beta_p x_p^*)}. \quad (20.10)$$

Since $\beta_1 \approx b_1, \beta_2 \approx b_2, \dots, \beta_p \approx b_p$, on the basis of (20.9), we have the estimate of the survival function for this specific individual

$$\hat{S}(t) \approx [S_0(t)]^{\exp(b_1 x_1^* + b_2 x_2^* + \dots + b_p x_p^*)}. \quad (20.11)$$

20.5 Computerized Experiments

Experiment 20.1 Kaplan–Meier estimate and log rank test The data come from Table 20.4. ID = 1 means failure to follow up, and ID = 0 means death. Line 23 is for Kaplan–Meier method, and line 24 is for log rank test.

Experiment 20.2 The estimation and test related to Cox model The data come from Table 20.5, and $Y = 0$ means failure to follow up, and $Y = 1$ means death.

20.6 Practice and Experiments

1. Calculate the two remission functions based on the data of Example 20.2 in Table 20.3 by Kaplan–Meier method and plot the two remission curves.
2. The relapse times of the patients with leukemia in the 6-MP treatment group and the control group are given in Table 20.8. Compare the two groups by log-rank test and Breslow test.

Program 20.1 Kaplan–Meier estimate and log rank test.

Line	Program	Line	Program
01	DATA LIFE;	14	24 1 1 34 1 2
02	INPUT T ID GROUP@@;	15	30 0 1 42 1 2
03	CARDS;	16	36 1 1 44 1 2
04	2 0 1 12 1 2	17	40 1 1 53 1 2
05	3 0 1 16 0 2	18	45 1 1 59 1 2
06	9 0 1 19 0 2	19	9 0 2 62 1 2
07	10 0 1 19 1 2	20	;
08	10 0 1 20 1 2	21	PROC PRINT;
09	12 1 1 20 1 2	22	PROC LIFETEST METHOD=PL;
10	15 0 1 24 1 2	23	TIME T*ID(1);
11	15 1 1 24 1 2	24	STRATA GROUP;
12	16 0 1 30 1 2	25	RUN;
13	18 1 1 31 1 2		

Table 20.8 The relapse times of patients with leukemia in two groups.

Group	Relapse time										
6-MP	6 17 ⁺	6 19 ⁺	6 20 ⁺	6 ⁺ 22	7 23	9 ⁺ 25 ⁺	10 32 ⁺	10 ⁺ 32 ⁺	11 ⁺ 34 ⁺	13 35 ⁺	16
Control	1 8	2 11	2 11	3 12	4 12	4 15	5 17	5 22	8 23	8	8

Program 20.2 Cox regression.

Line	Program	Line	Program
1	DATA a;	21	12 37 0 1 0 1 12
2	INPUT td v0 tr1 tr2 tr3 y day;	22	17 37 0 0 1 1 9
3	v01=0;v02=0;	23	14 29 0 0 1 1 12
4	td1=0;td2=0;	24	13 13 0 0 1 1 12
5	IF v0 > 18 and v0 < 30 THEN v01 = 1;	25	17 31 0 0 1 1 10
6	IF v0 >= 30 THEN v02 = 1;	26	;
7	IF td > 12 AND td < 17 THEN td1 = 1;	27	PROC PHREG;
8	IF td >= 17 THEN td2 = 1;	28	MODEL day*y(0)
9	CARDS;		= tr1 tr2 tr3 td1 td2 v01 v02;
10	19 25 0 0 0 1 8	29	PROC PHREG;
11	17 16 0 0 0 1 9	30	MODEL day*y(0)
			= tr1 tr2 tr3 td1 td2;
12	19 37 0 0 0 1 8	31	PROC PHREG;
13	16 19 0 0 0 1 8	32	MODEL day*y(0)
			= tr1 tr2 tr3 v01 v02;
14	14 25 1 0 0 0 18	33	PROC PHREG;
15	13 18 1 0 0 1 17	34	MODEL day*y(0)=tr1 tr2 tr3 v0;
16	16 25 1 0 0 1 14	35	PROC PHREG;
17	9 10 1 0 0 1 15	36	MODEL day*y(0)=tr1 tr2 tr3 td;
18	9 22 0 1 0 1 15	37	PROC PHREG;
19	10 25 0 1 0 1 11	38	MODEL day*y(0)=tr1 tr2 tr3;
20	14 25 0 1 0 1 13	39	RUN;

3. Work out Cox regressions with single covariate and multiple covariates for the data in Table 20.5.
4. For the data in Table 20.5, taking the variable *day* as dependent variable, the others as independent variables, work out a multiple linear regression, and compare the results with those obtained by Cox regression.
5. Keep all the data in Table 20.7 unchanged but let the terminal date change to May 30, 1989. Work out a Cox regression, and compare the results with those in Table 20.7.

(1st edn. Futian Luo, Jiqian Fang; 2nd edn. Zhang Jinxin, Jiqian Fang)

Chapter 21

Log-Linear Model for Contingency Table and Poisson Regression

Log-linear models have wide applications. This chapter focuses on their applications for analysis of contingency table and Poisson regression.

21.1 Log-Linear Models for Contingency Table

Example 21.1 There are two (one conventional and one experimental) treatments for a disease. Patients can be classified into success or failure according to the treatment results as well as severe or non-severe according to their original disease severity. A summary of data is listed in Table 21.1. We want to know which treatment is more effective.

If we ignore the original three-dimensional data structure and combine data into Table 21.2, we can use a χ^2 -test to evaluate the association between treatments and their results. The $\chi^2 = 2954.534$ with degrees of freedom 1. The p -value is less than 0.0001. Thus, we can conclude that the two treatments have different efficacy. Furthermore, we can use odds ratio to determine which treatment is a better choice. The odds ratio $OR = 1084 \times 6871 / (5300 \times 9130) = 0.1539$, which implies that the conventional treatment is more effective.

If we further examine Table 21.1 conditioning on disease severity, we will see that the experimental treatment is more effective than conventional one for non-severe patients with an $OR = 98 \times 5820 / (5 \times 5251) = 21.7239$. Similarly, it is more effective than conventional treatment for severe patients with an $OR = 986 \times 1051 / (9125 \times 49) = 2.3177$.

Therefore, we can see that it is incorrect to ignore disease severity. As there are 99% of patients having severe disease in the experimental treatment group and only 9% for the conventional treatment group, ignoring the

Table 21.1 A three-dimensional $2 \times 2 \times 2$ contingency table.

Treatments (X)	Treatments results (Y)	Disease severity (Z)	
		Non-severe	Severe
Experimental	Success	98	986
	Failure	5	9125
Conventional	Success	5251	49
	Failure	5820	1051

Table 21.2 Combining Table 21.1 by ignoring disease severity.

Treatments (X)	Treatment results (Y)	
	Success	Failure
Experimental	1084	9130
Conventional	5300	6871

disease severity introduce bias and cannot properly evaluate the treatment efficacy.

From this example, we learn that we cannot arbitrarily combine data from high dimensional contingency tables into two-dimensional contingency tables. While it is possible to evaluate relationship between treatment and efficacy from a high dimensional table conditioning on other variables, it will be difficult when the dimension is high. Each cell tends to have a small or even zero count of patients in high dimensional contingency tables. An effective method to deal with this kind of data is to use a log-linear model.

21.1.1 *Log-linear models for three-dimensional contingency tables*

Example 21.2 To study the age and gender distributions of Colles' fracture in different years, a retrospective review of medical records was conducted in Tianjing Hospital and summarized in Table 21.3. The goal was to determine the effect of year (X), gender (Y), and age (Z) on the fracture frequencies.

Table 21.3 Frequencies of Colles' fractures in 1980 and 1981*.

Year (X)	Gender (Y)	Age Z (year)								
		0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-
1980	M	7	82	113	35	37	39	38	13	5
	F	6	14	21	20	57	104	57	32	6
1981	M	19	121	174	57	49	57	31	17	6
	F	10	31	56	31	57	134	99	43	11

*Results from Cao Xiutang: Chinese Health Statistics, 1992, 9(5); 30

Table 21.4 Log-linear models and goodness-of-fit for data in Table 21.3.

Model structure	<i>df</i>	G^2	<i>P</i> -value	Pearson χ^2	<i>P</i> -value
1. Mutual Independence (<i>X</i> , <i>Y</i> , <i>Z</i>)	25	389.9990	<0.0001	367.2081	<0.0001
2. Partial independence (<i>X</i> , <i>Y</i> <i>Z</i>)	17	23.0782	0.1467	22.8258	0.1550
3. Partial independence (<i>Y</i> , <i>X</i> <i>Z</i>)	17	380.9824	<0.0001	359.7271	<0.0001
4. Partial independence (<i>Z</i> , <i>X</i> <i>Y</i>)	24	389.8811	<0.0001	367.8406	<0.0001
5. Conditional independence (<i>XY</i> , <i>XZ</i>)	16	380.8644	<0.0001	359.6620	<0.0001
6. Conditional independence (<i>XY</i> , <i>YZ</i>)	16	22.9602	0.1148	22.7309	0.1211
7. Conditional independence (<i>XZ</i> , <i>YZ</i>)	9	14.0615	0.1202	13.9477	0.1242
8. Homogeneous Association (<i>XY</i> , <i>XZ</i> , <i>YZ</i>)	8	11.7856	0.1610	11.7442	0.1630
9. Saturated (<i>XYZ</i>)	0	0.0000		0.0000	

Using PROC GENMOD in SAS, we can fit nine possible models for this $2 \times 2 \times 9$ three-dimensional contingency table (Table 21.3). There are several goodness-of-fit statistics for each model, including commonly used parameters of deviance, G^2 and Pearson χ^2 . We summarize the results in Table 21.4.

The first column of Table 21.4 is the model structure. Nine possible models are listed with the most complete model as the *saturated model*. When some of the terms in the saturated model are set as zeros, we have other possible models.

In an $I \times J \times K$ contingency table, let $\mu_{i,j,k}$ be the expected frequency in the cell of the i th level of X ($i = 1, 2, \dots, I$), the j th level of

$Y(j = 1, 2, \dots, J)$, and the k th level of $Z(k = 1, 2, \dots, K)$. In a saturated model XYZ , the expected frequency is decomposed as

$$\ln \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}. \quad (21.1)$$

Here, λ is the constant term. λ_i^X is the main effect of X at the i th level. The larger λ_i^X is, the higher is the corresponding expected frequency μ_{ijk} . To facilitate comparisons, $\lambda_I^X = 0$. Similarly, λ_j^Y and λ_k^Z are the corresponding main effects of Y and Z . λ_{ij}^{XY} , λ_{ik}^{XZ} and λ_{jk}^{YZ} are the first order interaction effects at the corresponding levels. Again, to facilitate comparisons, we make zero interactions when X is at the I th level or Y is at the J th level or Z is at the K th level. That is $\lambda_{Ij}^{XY} = \lambda_{iJ}^{XY} = \lambda_{Ik}^{XZ} = \lambda_{iK}^{XZ} = \lambda_{jK}^{YZ} = \lambda_{jK}^{YZ} = 0$. λ_{ijk}^{XYZ} is the second order interaction effects of X , Y and Z at levels of i , j and k . We again make $\lambda_{Ijk}^{XYZ} = \lambda_{iJk}^{XYZ} = \lambda_{ijk}^{XYZ} = 0$.

When we fit a contingency table to a saturated model, the expected frequencies in all cells should be identical to the observed frequencies. This is because we simply change the expression of the contingency table using a set of different parameters. Therefore, a saturated model has no practical use. If there is no simpler model than the saturated model, the data cannot be analyzed by a log-linear model.

Homogeneous Association Model: When all second order interaction effects are zero ($\lambda_{ijk}^{XYZ} = 0$), the model in (21.1) becomes a homogeneous association model and is denoted by (XY, XZ, YZ) . This notation indicates that the associations of any two variables at any given levels of the third variable are always the same.

Conditional Independent Model: If not only all second order interaction effects are zero ($\lambda_{ijk}^{XYZ} = 0$), but also the first order interaction effects between X and Y are zero ($\lambda_{ij}^{XY} = 0$), the model (21.1) becomes a conditional independent model denoted by (XZ, YZ) when not all the first order interaction effects between X and Z as well as Y and Z are zero ($\lambda_{ik}^{XZ} \neq 0$ and $\lambda_{jk}^{YZ} \neq 0$, for some i , j , and k). In such a model, X and Y associate with Z . However, for any given level of Z , X and Y are independent. Similar conditional independent models include (XY, XZ) and (XY, YZ) .

Partial Dependent Model: If the second order interaction effects and the first order interaction effects between X and Z as well as Y and Z are all zero ($\lambda_{ijk}^{XYZ} = \lambda_{ik}^{XZ} = \lambda_{jk}^{YZ} = 0$) but not all the interaction effects between

X and Y are zero ($\lambda_{ij}^{XY} \neq 0$, for some i and j), the model (21.1) becomes a partial independent model denoted by (Z, XY) . X is independent of Z and so is Y in this model. But X and Y depend on to each other. Other partial independent models are (X, YZ) and (Y, XZ) .

Mutual Independent Mode: The simplest model for three-dimensional contingency table is the mutual independent model denoted by (X, Y, Z) . In this model, there are only main effects and all interactions are zero. All variables are independent of each other.

The log-linear models are hierarchical, in the sense that when higher order interaction effects are contained in a model, the lower order interaction effects and the main effects are also contained. For example, a model that has λ_{ijk}^{XYZ} should include all first order interactions λ_{ij}^{XY} , λ_{ik}^{XZ} , λ_{jk}^{YZ} and the main effects of λ_i^X , λ_j^Y , and λ_k^Z . Therefore, the notation of XYZ can appropriately represent the model structure in (21.1). Another example is the model (XY, YZ) , which does not include the second order interaction and the first order interaction effects between Y and Z , and only has the main effects as well as the first order interactions between X and Y and the first order interactions between X and Z .

Columns 3 and 5 of Table 21.4 are statistics for the goodness-of-fit of the models. They are

$$G^2 = 2 \sum O_{ijk} \ln \frac{O_{ijk}}{E_{ijk}},$$

$$\text{Pearson's } \chi^2 = \sum \frac{(O_{ijk} - E_{ijk})^2}{E_{ijk}}.$$

Here, O_{ijk} are the observed frequencies in Table 21.3 and E_{ijk} are the expected frequencies under assumed models. The degrees of freedom of these two statistics are given in the second column.

df = total number of cells—number of independent parameters.

The numbers of independent parameters of these models are given in Table 21.5. For example, the total number of cells in Table 21.3 is $2 \times 2 \times 9 = 36$ and the number of degrees of freedom in a mutual independent model in Table 21.3 is $1 + (2 - 1) + (2 - 1) + (9 - 1) = 11$. Thus the corresponding number of degrees of freedom is 25. For a saturated model, the total number

Table 21.5 Degrees of freedom for models in an $I \times J \times K$ contingency table.

Parameters	Number of independent parameters
λ	1
λ_i^X	$I - 1$
λ_j^Y	$J - 1$
λ_k^Z	$K - 1$
λ_{ij}^{XY}	$(I - 1)(J - 1)$
λ_{ik}^{XZ}	$(I - 1)(K - 1)$
λ_{jk}^{YZ}	$(J - 1)(K - 1)$
λ_{ijk}^{XYZ}	$(I - 1)(J - 1)(K - 1)$
Total	IJK

of independent parameters is $I \times J \times K$, which is the same as the number of cells; thus, the number of degrees of freedom is 0.

In Table 21.4, when the p -values are relatively larger, the models fit the data better.

We can choose models based on the change in goodness-of-fit statistics between two hierarchical models, which is denoted by ΔG^2 and called deviance, and their difference in the degrees of freedom Δdf .

In Table 21.4, the deviance ΔG^2 between model (XZ, YZ) and (XY, XZ, YZ) is $14.0615 - 11.7856 = 2.2759$ and the difference in degrees of freedom is $\Delta df = 9 - 8 = 1$. The corresponding P -value for a $\chi^2_{(1)}$ distribution is 0.1314. Thus, the two models are not significantly different from each other. As the model (XZ, YZ) is simpler than (XY, XZ, YZ) , it is a better choice of the two. However, the deviance between models (X, YZ) and (XZ, YZ) is $\Delta G^2 = 23.0782 - 14.0615 = 9.0167$ and $\Delta df = 8$. The corresponding p -value is 0.3409, and because (X, YZ) is the simplest model, therefore, it is chosen as the best model. According to this model, the differences in frequencies of Colles' fracture between males and females (Y) were different for different age groups (Z), i.e., the interaction between age and gender was significant. For different years (X), the interactions were the same. Therefore, we can pool data across decades to reduce the problem to a two-dimensional contingency table.

Table 21.6 Estimated model parameters for (X, YZ) .

Parameter	Estimate	SE	Wald χ^2	P-value
λ	2.3121	0.2434	90.2552	0.0001
λ_1^X	-0.3799	0.0495	58.7858	0.0001
λ_1^Y	-0.4353	0.3870	1.2656	0.2606
λ_1^Z	-0.0606	0.3483	0.0303	0.8618
λ_2^Z	0.9734	0.2847	11.6922	0.0006
λ_3^Z	1.5106	0.2680	31.7765	0.0001
λ_4^Z	1.0986	0.2801	15.3886	0.0001
λ_5^Z	1.9030	0.2600	53.5739	0.0001
λ_6^Z	2.6391	0.2510	110.5054	0.0001
λ_7^Z	2.2166	0.2554	75.3215	0.0001
λ_8^Z	1.4843	0.2686	30.5317	0.0001
λ_{11}^{YZ}	0.9208	0.5007	3.3823	0.0659
λ_{12}^{YZ}	1.9419	0.4206	21.3184	<0.0001
λ_{13}^{YZ}	1.7510	0.4077	18.4471	<0.0001
λ_{14}^{YZ}	1.0253	0.4245	5.8332	0.0157
λ_{15}^{YZ}	0.1535	0.4125	0.1384	0.7098
λ_{16}^{YZ}	-0.4726	0.4054	1.3590	0.2437
λ_{17}^{YZ}	-0.3804	0.4131	0.8482	0.3571
λ_{18}^{YZ}	-0.4810	0.4432	1.1779	0.2778

To further illustrate the gender difference in Colles' fracture cases for different age groups, we can fit data to the model (X, YZ) and derive estimations of the 19 parameters in Table 21.6. As illustrated previously, we force $\lambda_2^X = \lambda_2^Y = \lambda_9^Z = 0$, and $\lambda_{21}^{YZ} = \lambda_{22}^{YZ} = \dots = \lambda_{29}^{YZ} = \lambda_{19}^{YZ} = 0$.

From Table 21.6, we can see that the p -value for λ_1^X is small, suggesting a significant difference in the total number of fractures between 1980 and 1981 (686 and 1003). The P -value for gender difference λ_1^Y is large, suggesting an insignificant difference in the total number fractures between males and females (900 and 789), P -values for $\lambda_2^Z, \lambda_3^Z, \lambda_4^Z, \lambda_5^Z, \lambda_6^Z, \lambda_7^Z, \lambda_8^Z$ are all small except λ_1^Z , suggesting a significant difference in total number of fractures between age groups (42, 248, 364, 143, 200, 334, 225, 105 and 28).

The P -values for the effects of interactions in Table 21.6 $\lambda_{11}^{YZ}, \lambda_{12}^{YZ}, \lambda_{13}^{YZ}, \lambda_{14}^{YZ}$ are small. This means that the gender differences in fracture numbers were significantly different among age groups 0–9, 10–19, 20–29 and 30–39 (male: 26, 203, 287 and 92; female: 16, 45, 77 and 51). Because of the positive interaction effects, the male had higher number of Colles' fracture than females in these age groups. Also, the effects of interactions $\lambda_{16}^{YZ}, \lambda_{17}^{YZ}, \lambda_{18}^{YZ}$ are negative, suggesting more female fractures than males in these age groups 50–59, 60–69 and 70–79 years. However, these differences did not reach statistical significance.

While goodness-of-fit statistics measure the global agreement between data and model, further understanding the differences between data and model in each cell of a contingency table is called *residual analysis*. To study the residuals of the data in Table 21.3 fitted by (X, YZ) , we use the procedure PROC GENMOD in SAS to obtain the output of the Pearson's residuals and standardized residuals in Table 21.7.

In Table 21.7, columns 3 and 7 are the observed frequencies (O) and the columns 4 and 8 are the expected frequencies (E) according to the model (X, YZ) . Columns 5 and 9 are the Pearson's residual for each cell in Table 21.3.

$$\text{Pearson's Residual} = \frac{O - E}{\sqrt{E}}.$$

When the expected frequency is large enough, Pearson's residual follows a normal distribution with zero mean. The Pearson's χ^2 statistics is the sum of all the squared Pearson's residuals. When one of the Pearson's residuals has its absolute value greater than 2, the observed frequency in the cell may be an outlier. If several cells have Pearson's residuals above 2 in absolute values, the model is likely inappropriate for the data. Contrary, if all Pearson's residuals of a model have absolute values below 2, this model is likely to be appropriate for the data.

In addition to Pearson's residuals, GENMOD in SAS also provides a standardized Pearson's residual. Its absolute value is usually larger than Pearson's residual. When the expected frequency is large, the standard residual follows a standard normal distribution with mean 0 and variance 1. Therefore, the standardized residual is preferred over the Pearson's residual. Similarly, we can use the absolute residual of 2 as the standard for outliers. When many of the standardize residuals have absolute values greater than 2,

Table 21.7 Expected frequencies and residuals for Table 21.3 fitted by (X, YZ) .

Gender	Age (year)	1980				1981			
		<i>O</i>	<i>E</i>	Pearson's residuals	Standard residuals	<i>O</i>	<i>E</i>	Pearson's residuals	Standard residuals
M	0-9	7	10.5601	-1.0955	-1.4327	19	15.4399	0.9060	1.4327
M	10-19	82	82.4500	-0.0496	-0.0686	121	120.5500	0.0410	0.0686
M	20-29	113	116.5672	-0.3304	-0.4706	174	170.4328	0.2732	0.4706
M	30-39	35	37.3665	-0.3871	-0.5166	57	54.6335	0.3202	0.5166
M	40-49	37	34.9295	0.3503	0.4666	49	51.0705	-0.2897	-0.4666
M	50-59	39	38.9911	0.0014	0.0019	57	57.0089	-0.0012	-0.0019
M	60-69	38	28.0249	1.8843	2.4967	31	40.9751	-1.5583	-2.4967
M	70-79	13	12.1847	0.2336	0.3058	17	17.8153	-0.1932	-0.3058
M	80-	5	4.4677	0.2518	0.3278	6	6.5323	-0.2083	-0.3278
F	0-9	6	6.4985	-0.1956	-0.2550	10	9.5015	0.1617	0.2550
F	10-19	14	18.2771	-1.0004	-1.3159	31	26.7229	0.8274	1.3159
F	20-29	21	31.2741	-1.8372	-2.4403	56	45.7259	1.5194	2.4403
F	30-39	20	20.7140	-0.1569	-0.2067	31	30.2860	0.1297	0.2067
F	40-49	57	46.3020	1.5722	2.1127	57	67.6980	-1.3002	-2.1127
F	50-59	104	96.6655	0.7460	1.0444	134	141.3345	-0.6169	-1.0444
F	60-69	57	63.3606	-0.7991	-1.0884	99	92.6394	0.6608	1.0884
F	70-79	32	30.4618	0.2787	0.3700	43	44.5382	-0.2305	-0.3700
F	80-	6	6.9047	-0.3443	-0.4490	11	10.0953	0.2847	0.4490

Table 21.8 An artificial example of a $2 \times 2 \times 2$ contingency table.

X	Y	Z	
		1	2
1	1	4	3
	2	6	36
2	1	6	70
	2	9	840

the model is inappropriate. Otherwise, the model fits data well. In Table 21.7, six out of 36 cells had standardized residuals between 2 to 2.5. Overall, the model fits the data well.

21.1.2 Combining contingency tables

Example 21.3 Table 21.8 is an artificial example of a three-dimensional $2 \times 2 \times 2$ table. Calculate the conditional odds ratios to determine whether the data follow a log-linear model and also provide plans to reduce the dimension by combining some variables.

Solution For $Z = 1$ and 2, the odds ratio of X and Y is

$$OR_{XY|Z=1} = \frac{4 \times 9}{6 \times 6} = 1, \quad OR_{XY|Z=2} = \frac{3 \times 840}{36 \times 70} = 1.$$

Conditional odds ratios $OR_{XY|Z=1} = OR_{XY|Z=2} = 1$ implies the two variables X and Y independent of each other given Z ($\lambda_{ij}^{XY} = 0$).

However,

$$OR_{YZ|X=1} = \frac{4 \times 36}{6 \times 3} = 8, \quad OR_{YZ|X=2} = \frac{6 \times 840}{9 \times 70} = 8,$$

$$OR_{XZ|Y=1} = \frac{4 \times 70}{6 \times 3} = 15.5556, \quad OR_{XZ|Y=2} = \frac{6 \times 840}{9 \times 36} = 15.5556.$$

The conditional odds ratio greater than 1 implies that given X , the two variables Y and Z correlated with each other; and given Y , X and Z correlated with each other. Therefore, the data should be fitted with a conditional independent model of (XZ, YZ) .

Table 21.9 Data of a $2 \times 2 \times 2$ table from a partial independent model.

<i>X</i>	<i>Y</i>	<i>Z</i>	
		1	2
1	1	9	3
	2	6	2
2	1	6	2
	2	54	18

For such models, we can combine tables of X and Z at different levels of Y to study the relationship between X and Z in a lower dimensional table. We can also combine tables of Y and Z over X to study the relationship between Y and Z . For example, combining tables of two Y levels, the odds ratio between X and Z is 15.5556; similarly, combining tables of two X levels, the odds ratio between Y and Z is 8. As Z is the condition for X being independent of Y , we should not combine the tables across levels of Z to study the relationship between X and Y .

Example 21.4 Table 21.9 is a dataset following a partial independent model. Through computing the conditional odds ratios to understand this kind of models and to provide plan to combine tables.

Conditional odds ratios are

$$OR_{XY|Z=1} = \frac{9 \times 54}{6 \times 6} = 13.5, \quad OR_{XY|Z=2} = \frac{3 \times 18}{2 \times 2} = 13.5,$$

$$OR_{YZ|X=1} = \frac{9 \times 2}{6 \times 3} = 1, \quad OR_{YZ|X=2} = \frac{6 \times 18}{54 \times 2} = 1,$$

$$OR_{XZ|Y=1} = \frac{9 \times 2}{6 \times 3} = 1, \quad OR_{XZ|Y=2} = \frac{6 \times 18}{54 \times 2} = 1.$$

Therefore, the correct model is (Z, XY) . We can combine the tables across two levels of Z to study the relationship between X and Y . Then, we have

$$OR_{XY|Z} = \frac{(9 + 3) \times (54 + 18)}{(6 + 2) \times (6 + 2)} = 13.5.$$

Tables 21.8 and 21.9 are the artificial examples to help understanding the conditional independent model and partial independent model. When we fit Table 21.8 to model (XZ, YZ) and Table 21.9 to model (Z, XY) , the goodness-of-fit statistics should be zero, indicating these models being perfectly fit the data. It can be verified using the same procedures in the computer experiments.

21.1.3 Model selection for higher dimensional contingency tables

In Table 21.4, we demonstrated that there are a total of nine possible models to fit the data. However, when there are four categorical variables to study, there are many more possible models. We now use the following four-dimensional contingency table to illustrate the procedures to find the best model.

Example 21.5 The data presented in Table 21.10 contained the information of a group of pregnant women and their new-born infants. There were four variables: X for smoking or no-smoking of the mother; Y for the use of a contraceptive medicine prior to the pregnancy; Z for normal or abnormal status of infants; and W for the age of mothers. The levels of these variables were $I(i = 1, \dots, I)$ for X , $J(j = 1, \dots, J)$ for

Table 21.10 Relationship between smoking, use of contraceptive and abnormality of infants.

Age (year) (W)	Smoking (X)	Prior use of contraceptive medicine (Y)	Infant status (Z)	
			Normal	Abnormal
≤ 29	Yes	Yes	204	58
		No	330	67
	No	Yes	1051	210
		No	1014	178
30~34	Yes	Yes	125	31
		No	180	42
	No	Yes	582	144
		No	489	85
≥ 35	Yes	Yes	35	20
		No	35	10
	No	Yes	158	53
		No	119	31

*Data from Li J and Shi B., Chinese Health Statistics, 1988, 5(6):33.

$Y, K (k = 1, \dots, K)$ for Z , and $L (l = 1, \dots, L)$ for W . The goal was to evaluate whether smoking and contraceptive medicine affect the chance of abnormal development of infants.

Solution This is a $2 \times 2 \times 2 \times 3$ contingency table. For such a high dimensional model, we can firstly fit a mutual independent model (X, Y, Z, W) . Using the computer package to calculate the goodness-of-fit test statistics $G^2 = 107.0507$, Pearson $\chi^2 = 110.5091$, $df = 18$, $P < 0.0001$. This model has too large of residuals and is not sufficient to describe the data.

As the second step, we try to fit for conditional independent model based on the first order interaction terms (XY, XZ, XW, YZ, YW, ZW) . Then, we have the goodness-of-fit test statistics $G^2 = 6.8836$ and Pearson $\chi^2 = 6.9821$ with $df = 9$, and the P -value is between 0.6492 and 0.6390. The model seems reasonable.

The question now is to select an optimal model between (X, Y, Z, W) and (XY, XZ, XW, YZ, YW, ZW) that fits the data better.

We can start from mutual independent model (X, Y, Z, W) and add one of the interactions XY, XZ, XW, YZ, YW and ZW to the model. The corresponding G^2 's are 57.0446, 101.6260, 105.8828, 97.1512, 91.8399, 89.4464, with degrees of freedoms 17, 17, 16, 17, 16, 16, respectively. Among them, adding XY has the largest change in $G^2 = 107.0507 - 57.0446 = 50.0161$, and the change in degrees of freedom is 1. The P -value is < 0.0001 . All these suggest that adding XY will result in the most improvement of model. The candidate model is now (W, Z, XY) .

Based on the last model, we can add one more of the interactions XZ, XW, YZ, WY, WZ to it. The corresponding G^2 are 51.6199, 55.8767, 47.1451, 41.8338, 39.4403, with degrees freedom 16, 15, 16, 15, 15, respectively. The maximal change in G^2 is due to the interaction WZ , which has $\Delta G^2 = 57.0446 - 39.4403 = 17.6043$, $\Delta df = 17 - 15 = 2$ and P -value = 0.0002. Therefore, the better model is (XY, WZ) .

Consequently, we can add the term of $WY (P = 0.0005)$

$$\Delta G^2 = 39.4403 - 24.2294 = 15.2109, \quad \Delta df = 15 - 13 = 2$$

and add the term of $YZ (P = 0.0033)$

$$\Delta G^2 = 15.6017, \quad \Delta df = 1$$

and add the term of XZ ($P = 0.0079$)

$$\Delta G^2 = 7.9066, \quad \Delta df = 1.$$

But, the term WX cannot be added because the gain in goodness-of-fit test statistics is insufficient.

Thus, the best model is (XY, XZ, YZ, YW, ZW) with the log-linear model expressed as

$$\begin{aligned} \log \mu_{ijkl} = & \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^W + \lambda_{ij}^{XY} \\ & + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{jl}^{YW} + \lambda_{kl}^{ZW}. \end{aligned} \quad (21.2)$$

Re-examining the model by excluding any of the parameters will result in a large change of goodness-of-fit. Therefore the final model is (21.2). The terms λ_{ik}^{XZ} , λ_{jk}^{YZ} , λ_{kl}^{ZW} in (21.2) represent that infant status (Z) depends on smoking (X), prior use of contraceptive medicine (Y), and age of pregnancy (W).

It is worthwhile to notice that by stratifying the data into six groups according to smoking and age of pregnancy, the χ^2 statistics for association of abnormality infant (Z) and prior use of contractive medicine (Y) are $\chi_1^2 = 2.842$ (for age ≤ 29 smoking mothers), $\chi_2^2 = 1.362$, $\chi_3^2 = 0.053$, $\chi_4^2 = 5.580$, $\chi_5^2 = 2.357$, $\chi_6^2 = 0.973$. This seems to suggest that the infant abnormality is not associated with mother's prior use of contraceptive medicine. However, considering the overall test effects of the interaction terms λ_{jk}^{YZ} , such association cannot be ignored because the model without this interaction has a large change in G^2 ($\Delta G^2 = 18.7032 - 8.5477 = 10.1555$, $\Delta df = 12 - 11 = 1$, $P = 0.0014$). Therefore, in the presence of other factors and their interactions, the influence of mother's prior use of contraceptive medicine cannot be ignored. Separate analysis in 6×2 tables result in smaller sample size for each analysis and therefore, reduce the power to derive significant conclusions. The log-linear model can use all observations in the contingency table and has a better statistical power.

21.2 Poisson Regression

Example 21.6 Using a retrospective cohort design to study the data of occupational exposure and 5-year survival, Yu *et al.* (Chinese Health Statistics, 1996; 13(1):6) collected all 9572 workers who worked from August 18,

1966 to December 31, 1991 in a factory of Hubei province. The total number of person-years was 114,488 years. There were 159 deaths in the cohort. The goal was to determine whether the occupational exposure and age affect the mortality rate.

The numerator of morality rate (the number of deaths) was small and the denominator (person years) was large. Therefore, the estimated overall mortality rates were small. If we assume the deaths among different people are independent, the number of deaths should follow a Poisson distribution (see Chap. 2). Poisson regression analyzes the number of events and the rate of the event in association with covariates under the assumption of Poisson distribution using a log-linear model. For example, let us assume two covariates X and Y , the model can be defined as

$$\ln P_{ij} = \ln \frac{\mu_{ij}}{n_{ij}} = \lambda + \lambda_i^X + \lambda_j^Y. \quad (21.3)$$

Here, the subscript ij indicates the cell of i th level of X and j th level of Y ; μ_{ij} is the expected frequency of deaths in the cell, and n_{ij} is the observed person-years. The expected mortality rate $P_{ij} = \mu_{ij}/n_{ij}$; λ is the constant term in the model; λ_i^X represents the effect of the i th level of X and λ_j^Y represents the effect of the j th level of Y . If there are interaction between X and Y , λ_{ij}^{XY} can be added to make a saturated model. Reorganize (21.3) can derive an alternative format of the model:

$$\ln \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \ln n_{ij}. \quad (21.4)$$

Except the $\ln n_{ij}$ from data in the right-hand side of equation, (21.4) is similar to (21.1). Therefore, we can use the same goodness-of-fit approaches in the previous sections to look for the optimal model that fits the data best.

Let $X(i = 1, 2, \dots, 5)$ and $Y(j = 1, 2)$ be age and occupational exposure in Table 21.11. Using GENMOD in SAS to fit (21.3) and other possible models, we have summarized goodness-of-fit results in Table 21.12.

Model 2 has X effect only and Model 4 contains both effects of X and Y . Both of them have P -values above 0.05, suggesting they fit the data well. Which model should be chosen? The deviance of two models were $\Delta G^2 = 7.8617 - 2.4927 = 5.3690$ with $\Delta df = 5 - 4 = 1$ and a P -value of 0.0205. There is a significant difference between the two models. Therefore,

Table 21.11 Deaths from all causes in a factory of Hubei province*.

Age (Year)	Unexposed			Exposed		
	Number of deaths	Person years	Mortality rate	Number of deaths	Person years	Mortality rate
<40	39	59141	0.000659	30	34995	0.000857
40~49	14	6621	0.002114	33	9241	0.003571
50~59	3	650	0.004615	25	3115	0.008026
60~69	0	54	0	12	595	0.020168
≥70	0	9	0	3	67	0.044776

*Data from Yu Z. *et al.*, Chinese Health Statistics, 1996; 13(1):6.

Table 21.12 Goodness-of-fit results for all possible Poisson regression models.

Model	df	G^2	P-value	Pearson χ^2	P-value
1. λ	9	167.6069	<0.0001	409.6269	<0.0001
2. $\lambda + \lambda_i^X$	5	7.8617	0.1640	6.2687	0.2809
3. $\lambda + \lambda_j^Y$	8	133.9757	<0.0001	258.9776	<0.0001
4. $\lambda + \lambda_i^X + \lambda_j^Y$	4	2.4927	0.6459	1.5422	0.8191
5. $\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$	0	0.0000		0.0000	

it is worth to have Model 4 with an additional parameter in comparison to Model 2.

To interpret the model parameters, we can calculate the relative risk (RR) of mortality between exposed and unexposed subjects using formula (21.3), which is

$$RR = \frac{P_{i2}}{P_{i1}} = \frac{\exp(\lambda + \lambda_i^X + \lambda_2^Y)}{\exp(\lambda + \lambda_i^X + \lambda_1^Y)} = \exp(\lambda_2^Y - \lambda_1^Y).$$

Similarly, we can derive relative risks for all age groups against a reference age group. For categorical variables, GENMOD in SAS forces the parameter of the last level of a categorical variable as zero and use the estimated parameters to calculate the conditional relative risks. For example, the occupational exposure has two levels. The program makes the parameter $\lambda_2^Y = 0$. Therefore, the $RR = \exp(-\lambda_1^Y)$.

We can express the categorical variable as a binary indicator, named "expose", with expose = 1 for exposed group and expose = 0 for unexposed

group. Also, we can define dummy variables for age groups, namely age 4, age 5 age 6, and age 7, for age groups 40–49, 50–59, 60–69, and ≥ 70 years, respectively; when all age dummy variables are 0, the subject age is < 40 years. We can reformat Model 4 in Table 21.12 as follows:

$$\ln P = \beta_0 + \beta_1 \exp ose + \beta_2 age\ 4 + \beta_3 age\ 5 + \beta_4 age\ 6 + \beta_5 age\ 7. \quad (21.5)$$

The regression parameters are the natural logarithm of the relative risks, i.e., $RR_k = \exp(\hat{\beta}_k)$ for $k = 1, 2, \dots, 5$.

Using GENMOD to fit (21.5), we have the estimated regression coefficients and corresponding RR values in Table 21.13.

Here, the Wald $\chi^2 = [\hat{\beta}/SE.(\hat{\beta})]^2$ follows a $\chi^2_{(1)}$ and can be used to test whether the corresponding $\hat{\beta}_k$ should be zero. Because all the P -values in Table 21.13 are less than 5%, all parameters make significant contributions to the model.

The relative risk of exposed to unexposed group $RR = 1.5$, suggests a 50% increase in mortality rate for exposed subjects. Compare to subjects younger than 40 years old, the subjects with increments in age of 10, 20, and 30 years have elevated mortality rates 3.71, 8.50 and 20.48 times of the rate of young people respectively. The predicted mortality rates and the standardized residuals are showed in Table 21.14. The absolute values of standardized residuals are less than one in all ten cells, indicating a successful fit of the Poisson regression model.

When the expected frequency, such as deaths, is large and P_{ij} is small, the Poisson distribution behaves very similar to a binomial distribution, and

Table 21.13 Estimated regression coefficients and RR .

Variables	Parameter	$\hat{\beta}$	$SE(\hat{\beta})$	Wald χ^2	P-value	RR		
						(95% confidence interval)		
	β_0	-7.3903	0.1469	2531.5753	<0.0001			
expose	β_1	0.4086	0.1787	5.2300	0.0222	1.50	1.06,	2.15
age 4	β_2	1.3110	0.1927	46.2929	<0.0001	3.71	2.53,	5.39
age 5	β_3	2.1401	0.2356	82.5225	<0.0001	8.50	5.29,	13.34
age 6	β_4	3.0195	0.3239	86.8824	<0.0001	20.48	10.41,	37.11
age 7	β_5	3.7902	0.5950	40.5564	<0.0001	44.27	10.93,	117.08

Table 21.14 Predicted mortality rates and residuals.

Age	Unexposed			Exposed		
	Observed mortality rate (%)	Predicted mortality rate (%)	Standardized residuals	Observed mortality rate (%)	Predicted mortality rate (%)	Standardized residuals
<40	0.66	0.62	0.8973	0.86	0.93	-0.8973
40-49	2.11	2.29	-0.4417	3.57	3.45	0.4417
50-59	4.62	5.25	-0.2492	8.03	7.89	0.2492
60-69	—	12.71	0.8596	20.17	19.02	-0.8596
≥70	—	27.64	0.5194	44.78	41.06	-0.5194

therefore, the relative risk and odds ratio are similar numerically.

$$\log(P_{ij}) \approx \log \frac{P_{ij}}{1 - P_{ij}} = \log \text{it}(P_{ij}).$$

In applications, the estimated regression coefficients of a logistic regression analysis can be used as approximations to regression coefficients of the Poisson regression analysis under these conditions. It is up to the readers to verify this using computer experiments.

21.3 Computerized Experiments

GENMOD in SAS can be used to establish the generalized linear models. The code for MODEL selection can specify many kinds of models through proper link functions according to the types of variables and assumptions on their distribution. For example, to continuous explanatory variables, the option “LINK = identify DIST = normal” deals with continuous variable with normal distribution in regression analysis (Chaps. 8 and 13). If the explanatory variables are categorical variables, this option specifies ANOVA models (Chaps. 11 and 14). If the explanatory variables are mixed with continuous and categorical variables, ANOCOVA will be performed. If the option is “LINK = logit DIST = binomial”, the program will perform a logistic regression analysis (Chap. 16). In this chapter, we use the option “LINK = log DIST = Poisson” to ask for log-linear regression analysis.

Experiment 21.1 Program 21.1 establishes a log-linear model for data in Table 21.3 using the model (X, YZ) . The “MODEL = POI” is the

Program 21.1 Log-linear model for the data in Table 21.3.

Line	Program	Line	Program
01	OPTIONS LS= 70	08	19 121 174 57 49 57 31 17 6
	PS = MAX NOCENTRE;		
02	DATA LOGLIN;	09	10 31 56 31 57 134 99 43 11
03	DO X=1 TO 2;DO Y=1	10	;
	TO 2;DO Z= 1 TO 9;		
04	INPUT F@@;	11	PROC GENMOD; CLASS X Y Z;
	OUTPUT;END;END;END;		
05	CARDS;	12	MODEL F= X Y Z /LINK=LOG
06	7 82 113 35 37 39 38 13 5	13	DIST=POI
07	6 14 21 20 57 104 57 32 6	14	OBSTATS RESIDUALS;
			RUN;

same as "MODEL=POISSON." By adding requests of "OBSTATS" and "RESIDUALS", the program will give outputs of observed frequencies and expected frequencies (Pred), Pearson residuals (Reschi), and standardized residuals (StReschi). To get goodness-of-fit statistics for models (X, Y, Z), (Y, XZ), (Z, XY), (XY, XZ), (XY, YZ), (YZ, XZ) and (XYZ) in Table 21.4, we only need to change the codes of line 12 from "XY|Z" to "X Y Z", "Y X|Z", to "Z X|Y", "X|Y X|Z", to "X|Y Y|Z", "X|Z Y|Z" to "X|Y|Z."

Note that CATMOD can also perform the analysis of log-linear models. However, the exported model parameters of CATMOD are different from GENMOD. This is because CATMOD specified that the sum of regression parameters should be zero for categorical variables and GENMOD specified the regression coefficient of the highest level of a categorical variable to be zero. Relative results of the two procedures, however, are the same.

Experiment 21.2 Analysis of high dimensional contingency tables

Program 21.2 performs a log-linear regression for the data in Table 21.10. Lines 01 to 10 read data of Table 21.10 into a SAS dataset MUTILOGL. Lines 11 to 20 use PROC GENMOD to find mutual independent model (X, Y, Z, W), first order conditional independent model (XY, XZ, XW, YZ, YW, ZW) and second order conditional independent model, (XYZ, XYW, XZW, YZW) respectively, and finally the log-linear model of (21.2). Similar codes can be used to fit other models.

Program 21.2 Selection of log-linear models for high dimensional contingency tables.

Line	Program	Line	Program
01	DATA MUTILOGL;	13	RUN;
02	DO W = 1 TO 3;DO X = 1 TO 2;	14	PROC GENMOD;
03	DO Y = 1 TO 2;DO Z = 1 TO 2;	15	CLASS X Y Z W ;
04	INPUT F@@;		MODEL F = X Y Z W
05	OUTPUT;END;END;END;END;	16	@2/ LINK=LOG
06	CARDS;		DIST=POI; RUN;
07	204 58 330 67 1051 210 1014 178	17	PROC GENMOD;
08	125 31 180 42 582 144 489 85	18	CLASS X Y Z W;
09	35 20 35 10 158 53 119 31	19	MODEL F= X Y X Z Y Z
10	;		Y W Z W/LINK=LOG
11	PROC GENMOD;	20	DIST=POI OBSTATS
12	CLASS X Y Z W;		RESIDUALS;
	MODEL F = X Y Z W/		RUN;
	LINK = LOG DIST = POI;		

Experiment 21.3 Poisson regression analysis Lines 01 to 12 of Program 21.3 read the data in Table 21.11 to the program. Line 03 derives an offset variable LNN. Lines 04 and 05 create dummy variables of AGE4, AGE5, AGE6 and AGE7. Lines 13 to 14 calculate goodness-of-fit for various models, including G^2 and Pearson's χ^2 statistics. To save the space, the program only the Model 4 in Table 21.12. Models 1, 2, 3 and 5 can be fitted by the same codes except changes "EXPOSE AGE" in line 14 to nothing, "EXPOSE", "AGE", and "EXPOSE|AGE." Lines 17 and 18 fit the model of (21.5) and obtain estimates in Table 21.13 as well as the expected mortality rates and residuals. Lines 17 and 18 use the offset variable LNN to derive predicted mortality rates. The results are the same as the output from Lines 15 and 16.

To obtain the results of logistic regression, the MODEL statement should be changed to "LINK=LOGIT DIST=BIN". Lines 19 to 21 are for the

Program 21.3 Poisson regression analysis for data in Table 21.3.

Line	Program	Line	Program
01	DATA POISSON;	14	MODEL Y/N=EXPOSE AGE7/ LINK=LOG;
02	INPUT AGE \$ EXPOSE N Y @@;		DIST=POI;RUN;
03	LNN=LOG(N);	15	PROC GENMOD;
04	AGE4=(AGE='40-49');	16	MODEL Y/N=EXPOSE AGE4-AGE7/LINK=LOG
05	AGE5=(AGE='50- 59');		DIST=POILRCI OBSTATS RESIDUALS; RUN;
06	AGE6=(AGE='60-69');	17	PROC GENMOD;
07	AGE7=(AGE='>= 70');		
08	CARDS;	18	MODEL Y=EXPOSE AGE4-AGE7/LINK=LOG
09	<40 0 59141 39 <40 1 34995 30		DIST=POILRCI OBSTATS RESIDUALS;
10	40-49 0 6621 14 40-49 1 9241 33		OFFSET=LNN; RUN;
11	50-59 0 650 3 50-59 1 3115 25	19	PROC GENMOD;
12	60-69 0 54 0 60-69 1 595 12	20	MODEL Y/N=EXPOSE AGE4-AGE7/ LINK=LOG DIST=POILRCI OBSTATS
13	>= 70 0 9 0 >= 70 1 67 3		RESIDUALS;
	;	21	RUN;
	PROC GENMOD; CLASS EXPOSE AGE		

logistic regression analysis. Because EXPOSE, AGE4 to AGE7 were all specified as 0–1 binary variables, there is no need to use CLASS option.

Adding LRCI behind MODEL statement gets the confidence intervals of the estimated regression coefficients. Using the natural exponential functions, we can derive the limits of confidence intervals for relative risks.

21.4 Practice and Experiments

1. The data in Table 21.15 were collected to evaluate the effectiveness of a traditional Chinese medicine regimen to treat chronic bronchitis. In this dataset, patients were grouped according to treatment results (X), smoking (Y), and duration of diseases (Z). Use a log-linear model to analyze this data.

Table 21.15 Data for evaluation of a traditional Chinese medicine for chronic bronchitis.

Treatment (X)	Smoking (Y)	Disease duration (Z)			
		≤5 years	6–10 years	11–20 years	≥21 years
Success	Yes	20	16	14	5
	No	29	23	16	6
Failure	Yes	16	14	20	12
	No	10	12	14	11

Table 21.16 The age specific mortality rates of esophageal cancers in two countries.

Age group	Country A		Country B	
	Population	Number of deaths	Population	Number of deaths
<40	2001839	12	2015117	25
40–49	251678	91	250480	125
50–59	206947	307	191204	344
60–69	143893	460	114355	371
≥70	90270	292	51670	170

2. What do parameters $\lambda_i^X, \lambda_j^Y, \lambda_k^Z, \lambda_{ij}^{XY}, \lambda_{ik}^{XZ}, \lambda_{jk}^{YZ},$ and λ_{ijk}^{XYZ} mean in a log-linear model? What are the relationship between the saturated model, the homogenous association model, the conditional independent model, the partial independent model, and the mutual independent model in a three-dimensional contingency table? How are they represented by the log-linear regression coefficients?
3. Table 21.16 presents the age specific mortality rates of esophageal cancers in two countries. First calculate their crude cause-specific death rates. Then calculate their age adjusted cause-specific death rates using Poisson regression models. What are the predicted numbers of deaths caused by esophageal cancers in these countries? Which country has higher death rate? Are the age adjusted death rates similar to the crude death rates? Is the difference between countries statistically significant? What are the relative risks (RR) for country A versus B?

Part III

Design and Analysis for Medical Research



Chapter 22

Multi-Factor Analysis of Variance

Experiments may be designed as a single-factor experiment or a multi-factor experiment according to the number of factors. The single-factor design has only one treatment factor with G levels ($G \geq 2$), such as completely random design, randomized complete-block design and Latin-square design (see Chap. 7). The multi-factor design has more than two factors, such as factorial design and split-plot design. One of the important issues in multi-factor design is to analyze the interaction effect, which is different from single-factor design. This chapter is to introduce several methods of analysis of variance used in multi-factor designs.

22.1 Factorial Experiments and Analysis of Variance

22.1.1 Introduction

The distinct characters of multi-factor design from the single factor design are that the G treatment groups are formed by the combinations of two or more than two factors, and each factor has at least two levels. When the G treatment groups are formed by all possible combinations of all levels of all factors, the design is called a complete factorial design. Taking a nutrition experiment as an example, supposing that factor A stands for protein in food, factor B stands for fattiness, and each factor has two levels, normal or absence, there will be $G = 2^2 = 4$ treatment groups in the experiment,

which is called 2^2 factorial design. The treatment groups can be laid out as follows:

Factor B ($J = 2$)	Factor A ($I = 2$)	
	Normal (a_1)	Absence (a_2)
Normal (b_1)	a_1b_1	a_2b_1
Absence (b_2)	a_1b_2	a_2b_2

If the complete factorial design is chosen, N experiment units should be randomly assigned into four groups, formed by two levels of A and B , $(a_1\ b_1)$, $(a_2\ b_1)$, $(a_1\ b_2)$ and $(a_2\ b_2)$. The experiment results can be shown as

$$x_{ijk},\ i = 1, 2,\ j = 1, 2,\ k = 1, 2, \dots, n_{ij}.$$
 (22.1)

To test the differences among the four groups, the method of ANOVA for completely randomized design introduced in Chap. 7 can be used, but for further analysis, ANOVA for a complete factorial design, introduced in this section, should be used, say analyzing the simple effect, main effect or interaction between factors etc.

22.1.1.1 Simple effect

Simple effects are the difference among different levels of the same factor when levels of other factors are fixed. Taking the result of 2^2 factorial experiment in Table 22.1 for example, when factor B is fixed at level 1, the simple effect of factor A is 2; when factor B is fixed at level 2, the simple effect of factor A is 8. Meanwhile, when factor A is fixed at level 1, the

Table 22.1 Example of 2^2 factorial experiment design (mean).

Factor A ($I = 2$)	Factor B ($J = 2$)		Mean effect $\bar{x}_{i..}$	Simple effect
	b_1	b_2		
a_1	$\bar{x}_{11.} = 30$	$\bar{x}_{12.} = 36$	$\bar{x}_{1..} = 33$	$\bar{x}_{12.} - \bar{x}_{11.} = 6$
a_2	$\bar{x}_{21.} = 32$	$\bar{x}_{22.} = 44$	$\bar{x}_{2..} = 38$	$\bar{x}_{22.} - \bar{x}_{21.} = 12$
Mean effect $\bar{x}_{.j.}$	$\bar{x}_{.1.} = 31$	$\bar{x}_{.2.} = 40$	$\bar{x} = 35.5$	
Simple effect	$\bar{x}_{21.} - \bar{x}_{11.} = 2$	$\bar{x}_{22.} - \bar{x}_{12.} = 8$		

simple effect of factor B is 6 and when factor A is fixed at level 2, the simple effect of factor B is 12.

22.1.1.2 Main effect

The main effect is the average of all simple effects over all levels of given factor. In a 2^2 factorial experiment, the main effects of factors A and B are

$$A = \frac{1}{2}[(\bar{x}_{22.} - \bar{x}_{12.}) + (\bar{x}_{21.} - \bar{x}_{11.})], \quad (22.2)$$

$$B = \frac{1}{2}[(\bar{x}_{22.} - \bar{x}_{21.}) + (\bar{x}_{12.} - \bar{x}_{11.})]. \quad (22.3)$$

In Table 22.1, $A = (8 + 2)/2 = 5$, $B = (12 + 6)/2 = 9$.

22.1.1.3 Interaction

When simple effects of factor A quite differ over the levels of factor B , then this difference is called interaction between the two factors; and when simple effects of factor B quite differ over the levels of factor A , then this difference is also called interaction between the two factors; it can be proved that the above two differences are equal. In a 2^2 factorial experiment, the interaction between the two factors A and B can be calculated as follows:

$$AB = \frac{1}{2}[(\bar{x}_{22.} - \bar{x}_{12.}) - (\bar{x}_{21.} - \bar{x}_{11.})] \quad (22.4a)$$

or

$$BA = \frac{1}{2}[(\bar{x}_{22.} - \bar{x}_{21.}) - (\bar{x}_{12.} - \bar{x}_{11.})] \quad (22.4b)$$

$$AB = BA.$$

Like the main effects, here $1/2$ means average on per-unit basis. For the data in Table 22.1,

$$AB = \frac{1}{2}[8 - 2] = 3, \quad BA = \frac{1}{2}[12 - 6] = 3.$$

The four means in Table 22.1 can be illustrated by Fig. 22.1, where the two un-parallel lines indicate the interaction of factors A and B . On the contrary, if the two lines are parallel, then there is no interaction between the two factors.

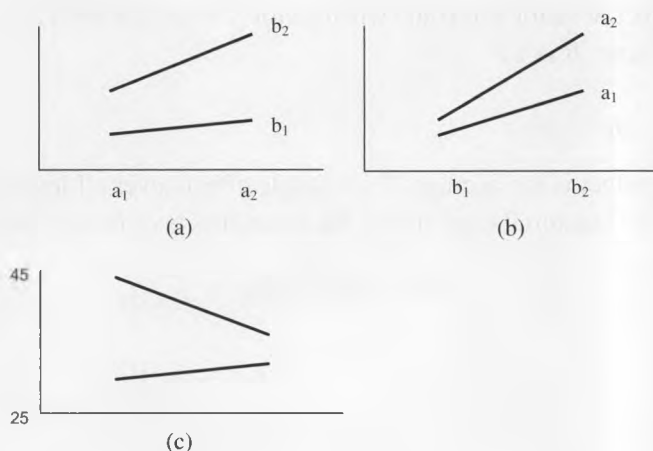


Fig. 22.1 Illustration of interaction of 2^2 factorial experiment.

If the statistical analysis indicates that the interaction exists, the simple effects of factors must be calculated separately. Otherwise, if there is no interaction, the responses to the two factors are independent of each other, then the main effect of a factor could be taken as the treatment effect of the factor. There are two factors being involved in the above-mentioned interaction so that it is called *two-factor interaction* or *first-order interaction*.

When the interactions between factors A and B quite differ over the levels of factor C , then we say the *three-factor interaction* or *second-order interaction* exists. It is much more difficult to illustrate the meaning of high-order interaction than the illustration in Fig. 22.1 for 2^2 factorial design.

Obviously, the factorial design can provide more information than the single factor design, especially reflecting the synergetic effect or antagonistic effect which is meaningful in medical researches on screening the best therapeutic regime, drug formula, experimental condition, etc. The disadvantage of factorial design is that when there are more factors (for instance more than three), the required number of treatments and experimental units will increase rapidly. Thus, if there are more factors being considered in an experiment, the orthogonal design is recommended to have initial screening first. Readers may refer to other books or references for the orthogonal design.

22.1.1.4 Experimental design

Taking all possible combinations of the levels over the factors to form treatment groups, the number of treatment groups is G . The most popular way is to assign the candidate experimental units into G treatment groups randomly. For instance, in a 2^2 factorial design the treatments can be regarded as a single factor treatment with four levels as follows:

$$\text{treatment 1} = (a_1 b_1),$$

$$\text{treatment 2} = (a_1 b_2),$$

$$\text{treatment 3} = (a_2 b_1),$$

$$\text{treatment 4} = (a_2 b_2).$$

However, it must be noticed that the factorial experiment with randomized complete design requires the numbers of units in different treatment groups to be all equal and more than two for replication; otherwise, the interaction among factors cannot be analyzed.

22.1.1.5 Analysis of variance

Let I and J represent the levels of factor A and factor B , $\bar{X}_{i..}$ ($i = 1, 2, \dots, I$) is the i th group mean of factor A , $\bar{X}_{.j}$ ($j = 1, 2, \dots, J$) is the j th group mean of factor B , and n_{ij} is the sample size of the treatment group with i th level of factor A and j th level of factor B . The total number of treatment groups is $G = IJ$, and \bar{X} represents the total mean. Then we define the sum of squared difference between all experiment results and the total mean as the Total SS

$$SS_{\text{total}} = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{X})^2 \quad (22.5)$$

and the weighted sum of squared difference between IJ of group means and the total mean as the treatment SS

$$SS_{\text{treatment}} = \sum_{i=1}^I \sum_{j=1}^J n_{ij} (\bar{X}_{ij} - \bar{X})^2 \quad (22.6)$$

and the difference between Total SS and Treatment SS as the Error SS

$$SS_E = SS_{\text{total}} - SS_{\text{treatment}}. \quad (22.7)$$

The treatment SS can also be decomposed as follows (to simplify thing, let $n_{ij} \equiv n_1$):

(1) SS for main effect of A , that is, the weighted sum of squared difference between mean of each level of factor A and the total mean

$$SS_A = \sum_{i=1}^I Jn(\bar{X}_{i..} - \bar{X})^2, \quad (22.8)$$

$$\nu_A = I - 1$$

where Jn is the number of units represented by the mean of each level of factor A .

(2) SS for main effect of B , that is, the weighted sum of squared difference between mean of each level of factor B and the total mean

$$SS_B = \sum_{j=1}^J In(\bar{X}_{.j.} - \bar{X})^2, \quad (22.9)$$

$$\nu_B = J - 1,$$

where In is the number of units represented by the mean of each level of factor B .

(3) SS for interaction effect between A and B , that is, the extra effect in addition to that of the main effects of A and B

$$SS_{AB} = SS_{\text{treatment}} - SS_A - SS_B,$$

$$\nu_{AB} = (I - 1)(J - 1).$$

Therefore,

$$SS_{\text{treatment}} = SS_A + SS_B + SS_{AB}, \quad (22.10)$$

$$\nu_{\text{treatment}} = \nu_A + \nu_B + \nu_{AB} = IJ - 1.$$

From (22.7),

$$SS_E = SS_{\text{total}} - SS_{\text{treatment}}. \quad (22.11)$$

The above equations are summarized in columns 1–3 of Table 22.2. The 4th column MS is the mean square of error obtained by SS/DF . The three

Table 22.2 Analysis of variance for two factors factorial experiment.

Source	<i>DF</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Between groups	$IJ - 1$	SS_{between}		
<i>A</i>	$I - 1$	SS_A	$MS_A = SS_A / (I - 1)$	$F_A = MS_A / MS_E$
<i>B</i>	$J - 1$	SS_B	$MS_B = SS_B / (J - 1)$	$F_B = MS_B / MS_E$
<i>AB</i>	$(I - 1)(J - 1)$	SS_{AB}	$MS_{AB} = SS_{AB} / (I - 1)(J - 1)$	$F_{AB} = MS_{AB} / MS_E$
Error	$IJ(n - 1)$	SS_E	$MS_E = SS_E / IJ(n - 1)$	
Total	$IJn - 1$	SS_{Total}		

Table 22.3 Passing rates of rabbit's axon after the operation of neural suture (%).

<i>B</i> (examination time after suture)	<i>A</i> (operation)			
	Adventitia suture (a_1)		Fasciculus suture (a_2)	
	1st month (b_1)	2nd month (b_2)	1st month (b_1)	2nd month (b_2)
	10	30	10	50
	10	30	20	50
	40	70	30	70
	50	60	50	60
	10	30	30	30
$\bar{x}_{i.}$ (treatment)	$\bar{x}_{11} = 24$	$\bar{x}_{12} = 44$	$\bar{x}_{21} = 28$	$\bar{x}_{22} = 52$
$\bar{x}_{i..}$ (factor <i>A</i>)	$\bar{x}_{1..} = 34$		$\bar{x}_{2..} = 40$	
$\bar{x}_{.j}$ (factor <i>B</i>)	$\bar{x}_{.1} = 26$		$\bar{x}_{.2} = 48$	

F values in the 5th column indicate whether the main effects *A*, *B* and the interaction *AB* are statistically significant.

The principle of decomposing *DF* and *SS* for factorial design with more than two factors is similar to that in Table 22.2, but it is very difficult to calculate by hand and better to use statistical software to create an ANOVA table.

Example 22.1 Analyze the result of 2^2 factorial experiment in Table 22.3, where 20 rabbits were used as the experimental units and randomly assigned into four groups. The treatment groups were formed according to all possible combinations of the method of the operation (factor *A*, two levels) and the examination time after suture (factor *B*, two levels).

Solution The example is 2^2 factorial experiment, $n = 5$.

$$SS_{\text{total}} = \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^5 (X_{ijk} - 37)^2 = 7420,$$

$$\begin{aligned} SS_{\text{treatment}} &= \sum_{i=1}^2 \sum_{j=1}^2 5 \times (\bar{X}_{ij.} - 37)^2 \\ &= 5 \times (24 - 37)^2 + 5 \times (44 - 37)^2 + 5 \times (28 - 37)^2 \\ &\quad + 5 \times (52 - 37)^2 = 2620, \end{aligned}$$

$$\begin{aligned} SS_A &= \sum_{i=1}^2 2 \times 5 \times (\bar{X}_{i..} - 37)^2 \\ &= 10 \times (34 - 37)^2 + 10 \times (40 - 37)^2 = 180, \end{aligned}$$

$$\begin{aligned} SS_B &= \sum_{j=1}^2 2 \times 5 \times (\bar{X}_{.j.} - 37)^2 \\ &= 10 \times (26 - 37)^2 + 10 \times (48 - 37)^2 = 2420, \end{aligned}$$

$$SS_{AB} = SS_{\text{treatment}} - SS_A - SS_B = 2620 - 180 - 2420 = 20,$$

$$SS_{\text{error}} = SS_{\text{total}} - SS_{\text{treatment}} = 7420 - 2620 = 4800.$$

The above results can be summarized into an ANOVA table for 2^2 factorial experiment (see Table 22.4).

The conclusion is that only main effect of time after suture, factor B , is statistically significant ($P < 0.05$). The main effect of factor B , sutured

Table 22.4 The ANOVA table for the data in Table 22.2.

Source	<i>DF</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>P</i>
Between groups	(3)	(2620)			
<i>A</i>	1	180	180	0.60	>0.05
<i>B</i>	1	2420	2420	8.07	<0.05
<i>AB</i>	1	20	20	0.07	>0.05
Error	16	4800	300		
Total	19	7420			

Table 22.5 Residual table for results of the factorial experiment with two factors.

A (suture methods)	Adventitia suture (a_1)		Fasciculus suture (a_2)	
	1st month (b_1)	2nd month (b_2)	1st month (b_1)	2nd month (b_2)
	-14	-14	-18	-2
	-14	-14	-8	-2
	16	-14	2	18
	26	26	22	8
	-14	16	2	-22
Treatment mean ($\bar{x}_{ij.}$)	24	44	28	52

after two months, is

$$B = \frac{1}{2}[(\bar{x}_{22.} - \bar{x}_{21.}) + (\bar{x}_{12.} - \bar{x}_{11.})] = \frac{1}{2}[(44 - 24) + (52 - 28)] = 22(\%).$$

Residual analysis is used to check the independence, normal distribution, and homogeneity of variances assumptions of analysis of variance model. The calculated residuals are listed in Table 22.5. The related residual plots can be drawn according to these data for check of the basic assumptions such as the independency, normality and homogeneity of variance for the observations.

If the above residual analysis fails to support the basic assumptions of ANOVA, then since the outcome measure of this experiment is the passing rates of rabbit's axon (%), ranging from 0 to 1, the arcsine of square root (formula (7.13) of Chap. 7) can be tried for variable transformation; after an ANOVA for the transferred data, the residual analysis can again be used to check the basic assumptions of independence, normality and homogeneity of variance for the transferred data.

22.2 Split-Plot Designs and Analysis of Variance

22.2.1 Introduction

To learn what the split-plot design is, let us show an example first.

Example 22.2 Ten domestic rabbits were randomly assigned into two groups, the rabbits in one group were injected with antitoxin drug (a_1),

Table 22.6 Diameters of skin-injured range of domesticated rabbits (mm).

Rabbit group (whole-plot)		Skin position (subplot)	
		Left	Right
I	1 (a_1)	(b_1)	(b_2)
	4 (a_1)	(b_2)	(b_1)
	6 (a_1)	(b_2)	(b_1)
	7 (a_1)	(b_1)	(b_2)
	10 (a_1)	(b_1)	(b_2)
II	2 (a_2)	(b_2)	(b_1)
	3 (a_2)	(b_2)	(b_1)
	5 (a_2)	(b_1)	(b_2)
	8 (a_2)	(b_1)	(b_2)
	9 (a_2)	(b_2)	(b_1)

those in another group were injected with saline as control (a_2). Then, the symmetric skin positions on two legs of each rabbit (denoted with L (left) and R (right)) were chosen to inject certain toxin with low dose (b_1) and with high dose (b_2) respectively. The diameters of the injuries on both legs of each rabbit were measured and the average injuries diameters were compared between the two groups, antitoxin drug versus saline. The design is given in Table 22.6, where the rabbit was the experimental unit for drugs (a_1 versus a_2), called *whole-plot*, and the skin area was the experimental unit for injury (b_1 versus b_2), called *subplots*. It seems that the subplots were "split" from a whole-plot so that such kind of design is called split-plot design.

What is the difference between split-plot design and factorial design? In fact, in factorial design, different levels of factor A and different levels of factor B all act on the same basic experiment units; but in split-plot design, different levels of factor A act on the whole-plots, while different levels of factor B only act on the subplots, a part of the whole-plot. The split-plot design is widely used in medical researches, especially in clinical medical studies. The flaw of split-plot design is that the efficiency of testing for main effect of factor A is usually lower than that for factor B .

22.2.2 Experimental design

In general, assume factor A is the factor with I levels a_1, a_2, \dots, a_I , applied to the whole-plots, factor B is the factor with J levels b_1, b_2, \dots, b_J , applied to the subplots, there are rI whole-plots ($r \geq 2$), each of which splits into J subplots.

The split-plot design can be a completely randomized design as well as a randomized complete block design according to whether a complete block is formed for the whole-plots or not. Now introduce the main steps of these two designs respectively.

22.2.2.1 Completely randomized split-plot design

Step 1 Randomly assign the rI whole-plots into I groups (with r whole-plots for each) which will receive the treatment a_1, a_2, \dots, a_I of factor A respectively.

Step 2 Randomly assign the J subplots of each whole-plot to receive the treatment b_1, b_2, \dots, b_J of factor B respectively.

In Table 22.6, $I = 2$, $J = 2$, $r = 5$. As the first step, rabbits 1, 4, 6, 7 and 10 were assigned antitoxin (treatment a_1), and the rest were assigned saline (treatment a_2). As the second step, the left positions of rabbits 1, 5, 7, 8 and 10 were assigned to receive the low dose toxin (treatment b_1) and their right positions were assigned to receive the high dose toxin (treatment b_2); the assignment for the rest of rabbits is just the opposite. Looking at the first two columns of Table 22.6, it is a completely random design for factor A ; while looking at the last two columns of the table, it is a random block design for factor B , where each rabbit is regarded as a block.

22.2.2.2 Randomized complete-block split-plot design

Step 1 Group all the whole-plots into r blocks, each of which contains I similar whole-plots.

Step 2 For each block, randomly assign the I whole-plots to I treatments of factor A respectively.

Step 3 For each whole-plot, randomly assign its J subplots to J treatments of factor B respectively.

Example 22.3 Suppose there were five litters of rabbits, from each of which, two rabbits with similar weight were picked out for the experiment.

Table 22.7 A randomized complete-block split-plot design for Example 22.3.

Litter (block)	Rabbit (whole-plot)	Skin position (subplot)	
		Left	Right
1	1 (a_1)	(b_1)	(b_2)
	2 (a_2)	(b_1)	(b_2)
2	3 (a_2)	(b_1)	(b_2)
	4 (a_1)	(b_1)	(b_2)
3	5 (a_2)	(b_1)	(b_2)
	6 (a_1)	(b_1)	(b_2)
4	7 (a_1)	(b_1)	(b_2)
	8 (a_2)	(b_1)	(b_2)
5	9 (a_2)	(b_1)	(b_2)
	10 (a_1)	(b_1)	(b_2)

For every two rabbits from the same litter, randomly assign one being injected with antitoxin, and another being injected with saline as control. For every rabbit, randomly select comparable skin positions on both legs (denoted with left and right) being injected with the low-dose and high-dose of toxin respectively.

This design is given in Table 22.7. As the first step, all the rabbits (whole-plots) are grouped into $r = 5$ blocks according to the litter of them, each block contains $I = 2$ rabbits; as the second step, every two rabbits in the same block are randomly assigned to antitoxin and saline (two treatments of factor A) respectively. As the third step, the two positions of each rabbit (subplots) are randomly assigned to low-dose and high-dose of toxin (two treatments of factor B) respectively. Looking at the first two columns of Table 22.7, it is a randomized complete block design for factor A ; while looking at the last two columns of the table; it is another random block design for factor B , where each rabbit is regarded as a block.

Comparing the design in Example 22.2 and that in Example 22.3, the only difference is the way of assigning the whole-plots to the treatments of factor A . In Example 22.2, it is a completely randomized design; while in Example 22.3, it is a randomized complete-block design. In fact, this

reflects the main difference between completely randomized split-plot design and randomized complete-block split-plot design.

22.2.3 Analysis of variance

The ANOVA table for the data of split-plot design consists of two parts. The first part is used to test the main effects of factor A ; the second part is used to test the main effect of factor B and interaction of A and B .

22.2.3.1 Completely randomized split-plot design

Assume that the experiment results are:

$$X_{ijk}, \quad i = 1, 2, \dots, I, \quad j = 1, 2, \dots, J, \quad k = 1, 2, \dots, n \quad (22.12)$$

of which, I is the total number of levels of factor A , J is the total number of levels of factor B , n is the repeated number for each level of factor A , and $I \times n$ is the total number of experiment units as the whole-plots.

(22.13)–(22.15) are the formulas related to the variance decomposition for the whole-plots:

- (1) Total SS for whole-plots, that is, the weighted sum of squared difference between the mean of whole-plot experiment unit ($\bar{X}_{i.k}$) and the total mean (\bar{X})

$$SS_{\text{whole-plots}} = \sum_{i=1}^I \sum_{k=1}^n J (\bar{X}_{i.k} - \bar{X})^2, \quad (22.13)$$

of which, J is the total number of subplot units represented by each mean of whole-plot experiment unit.

- (2) SS for the main effect of A , that is, the weighted sum of squared difference between the mean of each level of factor A ($\bar{X}_{i..}$) and the total mean

$$SS_A = \sum_{i=1}^I Jn (\bar{X}_{i..} - \bar{X})^2 \quad (22.14)$$

of which, Jn is the number of subplot units represented by means of each level of factor A .

Table 22.8 Analysis of variance for the whole-plots in completely randomized design.

Source	<i>DF</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
<i>A</i>	<i>I</i> - 1	<i>SS</i> _{<i>A</i>}	<i>MS</i> _{<i>A</i>} = <i>SS</i> _{<i>A</i>} / (<i>I</i> - 1)	<i>MS</i> _{<i>A</i>} / <i>MS</i> _{<i>E</i>1}
Error (<i>E</i> ₁)	<i>I</i> (<i>n</i> - 1)	<i>SS</i> _{<i>E</i>1}	<i>MS</i> _{<i>E</i>1} = <i>SS</i> _{<i>E</i>1} / <i>I</i> (<i>n</i> - 1)	
Whole-plots total (<i>T</i> ₁)	<i>In</i> - 1	<i>SS</i> _{whole-plots}		

(3) The *SS* for whole-plots error, that is, difference between the above two.

$$SS_{E1} = SS_{\text{whole-plots}} - SS_A. \quad (22.15)$$

The decomposition of *DF* and *SS* are listed in Table 22.8.

(22.16)–(22.19) are the formulas related to the variance decomposition for the subplots:

(1) Total *SS* for subplots, that is, the sum of squared difference between each observed value (*X*_{*ijk*}) and the total mean

$$SS_{\text{subplots}} = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^n (X_{ijk} - \bar{X})^2 \quad (22.16)$$

of which, *J* is the number of subplot units represented by means of whole-plots experiment units.

(2) *SS* for the main effect of *B*, that is, the weighted sum of squared difference between the mean of each level of factor *B* and the total mean

$$SS_B = \sum_{j=1}^J In(\bar{X}_{.j} - \bar{X})^2 \quad (22.17)$$

of which, *I*_{*n*} is the number of subplot units represented by means of each level of factor *B*.

(3) *SS* for effect of *A* and *B*, that is, the weighted sum of squared difference between the mean of treatment group and the total mean

$$SS_{A \text{ and } B} = \sum_{i=1}^I \sum_{j=1}^J n(\bar{X}_{ij.} - \bar{X})^2$$

of which, *n* is the number of whole-plots units represented by means of treatment groups.

Table 22.9 Decomposition of DF and SS for the subplots.

Source	DF	SS	MS	F
<i>B</i>	$J - 1$	SS_B	$MS_B = SS_B / (J - 1)$	MS_B / MS_{E_2}
<i>AB</i>	$(I - 1)(J - 1)$	SS_{AB}	$MS_{AB} = SS_{AB} / [(I - 1)(J - 1)]$	MS_{AB} / MS_{E_2}
Subplots error (E_2)	$I(r - 1)(J - 1)$	SS_{E_2}	$MS_{E_2} = SS_{E_2} / [I(r - 1)(J - 1)]$	
Subplots total	$rIJ - 1$	SS_{subplot}		

(4) *SS* for interaction of *A* and *B*, that is, the extra effect in addition to that of the main effects of *A* and *B*

$$SS_{AB} = SS_{A \text{ and } B} - SS_A - SS_B. \quad (22.18)$$

(5) *SS* for sub plots error

$$SS_{E_2} = SS_{\text{subplots}} - SS_{\text{whole-plots}} - SS_B - SS_{AB}. \quad (22.19)$$

The decomposition methods of *DF* and *SS* of the second part are listed in Table 22.9.

Example 22.2 (Cont'd) The data obtained under the design in Example 22.2 are showed in Table 22.10. The measurements were the diameters (mm) of skin-injured range. Analyze the data by ANOVA.

Solution The example is completely randomized split-plot design. The rabbits are the whole-plot units and the injection sites of the rabbits are the subplots. So there are ten whole-plot units and 20 subplot units.

(1) Analysis of variance for the first part

$$\begin{aligned}
 SS_{\text{whole-plots}} &= \sum_{i=1}^2 \sum_{k=1}^5 2 \times (\bar{X}_{i.k} - 19.68)^2 \\
 &= 2 \times (17.38 - 19.68)^2 + 2 \times (20.25 - 19.68)^2 \\
 &\quad + 2 \times (18.13 - 19.68)^2 + 2 \times (21.63 - 19.68)^2 \\
 &\quad + 2 \times (17.00 - 19.68)^2 + 2 \times (23.13 - 19.68)^2
 \end{aligned}$$

Table 22.10 Diameters of skin-injured range of domesticated rabbits (mm).

Injected drugs (<i>A</i>)	No. of rabbits	Toxin strength (<i>B</i>)				Total (<i>U_k</i>)	
		Low strength (<i>b₁</i>)		High strength (<i>b₂</i>)			
Antitoxin (<i>a₁</i>)	1	15.75(<i>L</i>)	80.25	19.00(<i>R</i>)	98.75	34.75	179.0
	4	15.50(<i>R</i>)	(<i>T₁</i>)	20.75(<i>L</i>)	(<i>T₂</i>)	36.25	(<i>A₁</i>)
	6	15.50(<i>R</i>)		18.50(<i>L</i>)		34.00	
	7	17.00(<i>L</i>)		20.50(<i>R</i>)		37.50	
	10	16.50(<i>L</i>)		20.00(<i>R</i>)		36.50	
Saline (<i>a₂</i>)	2	18.25(<i>R</i>)	98.75	22.25(<i>L</i>)	115.75	40.50	214.5
	3	18.50(<i>R</i>)	(<i>T₃</i>)	21.50(<i>L</i>)	(<i>T₄</i>)	40.00	(<i>A₂</i>)
	5	19.75(<i>L</i>)		23.50(<i>R</i>)		43.25	
	8	21.50(<i>L</i>)		24.75(<i>R</i>)		46.25	
	9	20.75(<i>R</i>)		23.75(<i>L</i>)		44.50	
Total	10	179.00(<i>B₁</i>)		214.50(<i>B₂</i>)		393.50	

$$\begin{aligned} &+ 2 \times (18.75 - 19.68)^2 + 2 \times (22.25 - 19.68)^2 \\ &+ 2 \times (18.25 - 19.68)^2 + 2 \times (17.38 - 19.68)^2 \\ &= 81.0125, \\ SS_A &= \sum_{i=1}^I Jn(\bar{X}_{i..} - 19.68)^2 \\ &= 2 \times 5 \times (17.90 - 19.68)^2 + 2 \times 5 \times (21.45 - 19.68)^2 \\ &= 63.0125, \\ SS_{E1} &= SS_{\text{whole-plots}} - SS_A = 81.0125 - 63.0125 = 18.0000. \end{aligned}$$

From Table 22.8 (or using statistical software), results of *DF*, *MS* and *F* values are listed in the first part of Table 22.11.

(2) Analysis of variance for the second part

$$\begin{aligned} SS_{\text{subplots}} &= \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^5 (X_{ijk} - 19.68)^2 = 146.1375, \\ SS_B &= \sum_{j=1}^2 2 \times 5 \times (\bar{X}_{.j.} - 19.68)^2 \end{aligned}$$

Table 22.11 The table of ANOVA for the data in Table 22.26.

Source	<i>DF</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>P</i>
Injected drugs (<i>A</i>)	1	63.0125	63.0125	28.01	<0.01
Rabbits' error (within whole-plots)	8	18.0000	2.2500		
Rabbits' total (whole-plots total)	(9)	(81.0125)			
Toxin dosage (<i>B</i>)	1	63.0125	63.0125	252.05	<0.01
<i>AB</i>	1	0.1125	0.1125	0.45	>0.05
Skin-place' error (within subplots)	8	2.0000	0.2500		
Skin-place' total (subplots total)	(10)	(65.1250)			
Total	19	146.1375			

$$= 2 \times 5 \times (17.90 - 19.68)^2 \\ + 2 \times 5 \times (21.45 - 19.68)^2$$

$$= 63.0125,$$

$$SS_{AB} = \sum_{i=1}^2 \sum_{j=1}^2 5 \times (\bar{X}_{ij.} - 19.68)^2 - SS_A - SS_B$$

$$= 5 \times (16.05 - 19.68)^2 + 5 \times (19.75 - 19.68)^2 \\ + 5 \times (19.75 - 19.68)^2 + 5 \times (23.15 - 19.68)^2 \\ - 63.0125 - 63.0125 = 0.1125,$$

$$SS_{E2} = SS_{\text{subplots}} - SS_{\text{whole-plots}} - SS_B - SS_{AB} \\ = 146.1375 - 81.0125 - 63.0125 - 0.1125 = 2.0000.$$

From Table 22.9 (or using statistical software), results of *DF*, *MS* and *F* values are listed in the second part of Table 22.11.

The conclusion is that no statistical significance was found for interaction effect between *A* and *B*. The main effects of drugs (*A*) and toxin dosage (*B*) are statistically significant to the diameters of skin-injured range ($P < 0.01$). The antitoxin might be able to protect the rabbits' skin from the injuries with mean diameter of 17.9 mm, which decreases 3.6 mm compared to the control group of 21.5 mm.

Table 22.12 ANOVA for the whole-plots in randomized complete-block design.

Source	<i>DF</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Between blocks	$n - 1$	SS_{B1}	$MS_{B1} = SS_{B1}/(n - 1)$	MS_{B1}/MS_{E1}
A	$I - 1$	SS_A	$MS_A = SS_A/(I - 1)$	MS_A/MS_{E1}
Whole-plots	$I(n - 1)$	SS_{E1}	$MS_{E1} = SS_{E1}/(r - 1)(J - 1)$	
Error (E_1)				
Whole-plots total (T_1)	$In - 1$	$SS_{\text{whole-plots}}$		

22.2.3.2 Randomized complete-block split-plot design

To randomized complete-block split-plot design, the repeated number (n) of factor A in (22.12) is the block number. An SS for whole-plot blocks (22.20) is added to the analysis of variance of the first part of whole-plots units in Table 22.8, and the first part is shown in Table 22.12. The decomposition of variance for subplots in the second part of analysis of variance is the same as that of completely randomized split-plot design (22.16)–(22.19), and the table of analysis of variance is the same as in Table 22.9.

(1) SS for whole-plot blocks, that is, the weighted sum of square of difference between means of each block and total mean

$$SS_{\text{whole-plot blocks}} = \sum_{k=1}^n IJ (\bar{X}_{..k} - \bar{X})^2 \tag{22.20}$$

of which, IJ is the number of subplots units represented by means of each block.

Example 22.3 (Cont'd) Suppose an experiment was carried out under the design in Example 22.3, and the data were shown in Table 22.13. Analyze the data by ANOVA.

Solution As known before, it is a randomized completely complete-block split-plot design. There are five blocks, each block has two rabbits; the rabbits are whole-plots with two injection places for each as the subplots. Block number $n = 5$, treatment number, $I = 2$, $J = 2$. Descriptive statistics are calculated for experiment results in Table 22.13, and the statistics to show the whole-plots variance include rabbits mean, blocks mean (Table 22.13); main effect for factor A , main effect for factor B , group means of four

Table 22.13 Diameters of skin-injured range (mm).

Blocks	Rabbits (whole-plot unit)	Skin positions (subplot unit)		Rabbit mean	Block mean
		Left	Right		
1	1	15.75($a_1\ b_1$)	19.00($a_1\ b_2$)	17.38	18.81
	2	18.25($a_2\ b_1$)	22.25($a_2\ b_2$)	20.25	
2	3	18.50($a_2\ b_1$)	21.50 ($a_2\ b_2$)	20.00	19.06
	4	15.50($a_1\ b_1$)	20.75($a_1\ b_2$)	18.13	
3	5	19.75($a_2\ b_1$)	23.50 ($a_2\ b_2$)	21.63	19.31
	6	15.50($a_1\ b_1$)	18.50 ($a_1\ b_2$)	17.00	
4	7	17.00($a_1\ b_1$)	20.50 ($a_1\ b_2$)	18.75	20.94
	8	21.50($a_2\ b_1$)	24.75 ($a_2\ b_2$)	23.13	
5	9	20.75 ($a_2\ b_1$)	23.75 ($a_2\ b_2$)	22.25	20.25
	10	16.50 ($a_1\ b_1$)	20.00 ($a_1\ b_2$)	18.25	
<hr/>					
$\bar{x}_{11} = 16.05$	$\bar{x}_{12} = 19.75$	$\bar{x}_{1.} = 17.9$	$\bar{x}_{2.} = 21.5$	$\bar{x}_{1..} = 17.9$	$\bar{x}_{2..} = 21.5$
$\bar{x}_{21} = 19.75$	$\bar{x}_{22} = 23.15$			$\bar{x} = 19.68$	

Table 22.14 The first two columns of the ANOVA Table.

Source	DF
Injected drugs (A)	1
Between blocks	4
Rabbits error (within whole-plots)	4
Rabbits total (whole-plots total)	(9)
Toxin dosage (B)	1
AB	1
Skin-place error (within subplots)	8
Skin-place total (subplots total)	(10)
Total	19

treatment groups. Then the table of analysis of variance can be obtained from Tables 22.12 and 22.9 (or from statistical software). Compared to Table 22.8, a block variance is added to the whole-plots SS , meanwhile, a DF of 4 is lost (see Table 22.14). If the block SS is big which can decrease whole-plots error to increase the power of factor A . Otherwise, the block SS can be combined to whole-plots error.

22.3 Cross-Over Design and Analysis of Variance

22.3.1 Introduction

Example 22.4 In order to detect the effect of a sedative, a clinical trial is designed such that the objects are insomnia patients in a clinic. A placebo denoted as a_1 is taken as control, the treatment with the sedative denoted as a_2 . Assuming the patients could receive a_1 and a_2 respectively during two periods of time, and the patients are randomly assigned to the two groups, a cross-over design can be laid out as follows:

Treatment group	Two periods of time		
	I	Washout	II
1st group	a_1	No treatment	a_2
2nd group	a_2	No treatment	a_1

A notable characteristic of cross-over design in a clinical trial is that the experimental time can be divided into two periods; the objects are able to receive two treatments at different periods with different sequence. Some patients receive the treatments in the order a_1 and a_2 , others with the opposite order a_2 and a_1 . Though the treatment of cross-over design is a single factor treatment, another two factors may also influence the result of experiment. They are the sequences of treatment and the phases of periods. Therefore, cross-over design is actually a multi-factor experiment with three factors: the treatment, expressed as factor A ; the sequences of treatment, expressed as factor B ; and the phases of periods, expressed as factor C . However, the main purpose of cross-over design is to test the main effect of factor A under the assumption that there is no interaction among the three factors A , B and C . As the difference of treatments in cross-over trial is compared within objects, the variation among individuals can be avoided. It is especially suitable to the clinical trail to control the difference among individuals. A strict requirement is that the treatment effect in phase I would not transit to phase II, that is to say, there is no carry-over effect when a treatment stop, or the carry-over effects of both treatments are equal. To ensure this assumption, some "washout periods" may be set up before a treatment, and between two phases of treatment. During the washout periods, all drugs or treatments should be stopped to wait for the carry-over effect to disappear.

In clinical trials, cross-over design is usually used in comparing the effect of a drug or a therapeutic with the routine one for remission of symptoms, for instance, relieving pain, sedation, reducing blood pressure and resisting rheumatism, etc.

22.3.2 Two phase cross-over design

The design is quite simple for the cross-over design with only two treatments, a_1 and a_2 , two phases of periods, I and II, which called two phase cross-over design. Just randomly assign N objects into two treatment groups, the individuals in the first group receive the treatment a_1 in phase I and a_2 in phase II, those in the second group receive a_2 in phase I and a_1 in phase II. Notice that it is better to keep two treatment groups with the same number of cases during randomization for making the statistical analysis simpler.

22.3.3 Analysis of variance

To assume i as subject, j as experiment phase, the experiment result of the k treatment is

$$x_{ijk}, \quad i = 1, 2, \dots, N, \quad j = 1, 2, \quad k = 1, 2. \quad (22.21)$$

The experiment results of completely randomization two phases cross-over design are listed in Table 22.15.

Similarly, we have to calculate a series of "Sum of Squares" for analysis of variance:

- (1) SS for total, that is, the sum of squared differences between all observed values and the total mean

$$SS_{\text{total}} = \sum_{i=1}^n \sum_{j=1}^2 \sum_{\substack{k=1 \\ k \neq j}}^2 (X_{ijk} - \bar{X})^2 + \sum_{i=n+1}^N \sum_{j=1}^2 \sum_{\substack{k=1 \\ k \neq j}}^2 (X_{ijk} - \bar{X})^2. \quad (22.22)$$

- (2) SS for subjects, that is, the weighted sum of squared differences between the subject means ($\bar{X}_{i..}$) and the total mean

$$SS_{\text{subject}} = \sum_{i=1}^N 2(\bar{X}_{i..} - \bar{X})^2, \quad (22.23)$$

Table 22.15 Experiment results of two phases cross-over design.

Treatment order	Subject number after randomization	Experiment phases		Subject mean($\bar{x}_{i..}$)
		I	II	
A	1	X_{111}	X_{122}	$\bar{x}_{1..}$
	2	X_{211}	X_{222}	$\bar{x}_{2..}$
	\vdots	\vdots	\vdots	\vdots
	\vdots	\vdots	\vdots	\vdots
	n	X_{n11}	X_{n22}	$\bar{x}_{n..}$
		$\bar{X}_{.11}$	$\bar{X}_{.22}$	
B	$n+1$	$X_{n+1,12}$	$X_{n+1,21}$	$\bar{x}_{n+1..}$
	$n+2$	$X_{n+2,12}$	$X_{n+2,21}$	$\bar{x}_{n+2..}$
	\vdots	\vdots	\vdots	\vdots
	\vdots	\vdots	\vdots	\vdots
	N	$X_{N,12}$	X_{N21}	$\bar{x}_{N..}$
		$\bar{X}_{.12}$	$\bar{X}_{.21}$	
Phase mean ($\bar{x}_{.j.}$)		$\bar{x}_{.1.}$	$\bar{x}_{.2.}$	
Treatment mean ($\bar{x}_{..k}$)		$\bar{x}_{..1}$	$\bar{x}_{..2}$	
Total mean		$\bar{x}_{...}$		

where 2 is the number of the observed values represented by each subject mean.

- (3) SS for experiment phase, that is, the weighted sum of squared difference between the mean of two phases ($\bar{X}_{.j.}$) and the total mean.

$$SS_{\text{phase}} = \sum_{j=1}^2 N(\bar{X}_{.j.} - \bar{X})^2, \quad (22.24)$$

where N is the number of the observed values represented by each phase mean.

- (4) SS for treatment, that is, the weighted sum of squared difference between the means of two treatment ($\bar{X}_{..k}$) and the total mean

$$SS_{\text{treatment}} = \sum_{k=1}^2 N(\bar{X}_{..k} - \bar{X})^2, \quad (22.25)$$

Table 22.16 Analysis of variance for two phases cross-over design.

Source	DF	SS	MS	F
Patients	$N - 1$	SS_{subject}	$MS_{\text{subject}} = SS_{\text{subject}} / (N - 1)$	$MS_{\text{subject}} / MS_E$
Phases	1	SS_{phase}	$MS_{\text{phase}} = SS_{\text{phase}} / 1$	MS_{phase} / MS_E
Treatments	1	$SS_{\text{Treatment}}$	$MS_{\text{Treatment}} = SS_{\text{Treatment}} / 1$	$MS_{\text{Treatment}} / MS_E$
Error	$N - 2$	SS_E	$MS_E = SS_E / (N - 2)$	
Total	$2N - 1$	SS_{Total}		

where N is the number of the observed values represented by each treatment mean.

(5) SS for Error

$$SS_E = SS_{\text{total}} - SS_{\text{subject}} - SS_{\text{phase}} - SS_{\text{treatment}}. \quad (22.26)$$

Put the above results to Table 22.16 to calculate and complete analysis of variance for two phases cross-over design.

Example 22.5 To compare the effects of two drugs a_1 and a_2 on losing weight, through a two-phase cross-over design, 12 objects with obesity were randomly divided into two groups, the first group used drug a_1 in phase I and drug a_2 in phase II; the second group used drug a_2 in phase I and drug a_1 in phase II. Phases I and II lasted four weeks respectively. The results of observation are listed in Table 22.17, analyze the data.

Solution Calculate descriptive statistics, including subjects mean, phase mean and treatment mean (Table 22.33). Put them into the formulas (22.22)–(22.26).

$$\begin{aligned}
 SS_{\text{total}} &= \sum_{i=1}^{12} \sum_{j=1}^2 (X_{ij} - 2.342)^2 = 109.7462 \quad SS_{\text{total}} \\
 &= \sum_{i=1}^{12} \sum_{j=1}^2 (X_{ij} - 2.342)^2 = 109.7462,
 \end{aligned}$$

Table 22.17 Losing weights (kg).

Drug sequence	No. of subjects	Phase I	Phase II	Subject mean $\bar{x}_{i..}$
1st group ($a_1 \rightarrow a_2$)	1	6.129	-0.454	2.838
	2	2.497	0.908	1.703
	3	4.313	0.454	2.384
	4	4.540	2.724	3.632
	5	1.498	1.135	1.317
	6	8.172	4.313	6.243
2nd group ($a_2 \rightarrow a_1$)	7	4.449	2.043	3.246
	8	4.994	1.816	3.405
	9	0.454	0.136	0.295
	10	0.227	1.271	0.749
	11	1.589	1.271	1.430
	12	0.136	1.589	0.863
Phase mean $\bar{x}_{.j.}$		3.250	1.434	2.342 (\bar{x})
Treatment mean $\bar{x}_{.k}$		2.940	1.744	

$$\begin{aligned}
 SS_{\text{subject}} &= \sum_{i=1}^{12} 2 \times (\bar{X}_{i..} - 2.342)^2 \\
 &= 2 \times (2.838 - 2.342)^2 + 2 \times (1.703 - 2.342)^2 \\
 &\quad + 2 \times (2.384 - 2.342)^2 \\
 &\quad + \cdots + 2 \times (1.430 - 2.342)^2 \\
 &\quad + 2 \times (0.863 - 2.342)^2 = 60.5631,
 \end{aligned}$$

$$\begin{aligned}
 SS_{\text{phase}} &= \sum_{j=1}^2 N(\bar{X}_{.j.} - \bar{X})^2 \\
 &= 12(3.250 - 2.342)^2 + 12(1.434 - 2.342)^2 = 19.7871,
 \end{aligned}$$

$$\begin{aligned}
 SS_{\text{treatment}} &= \sum_{k=1}^2 12(\bar{X}_{.k} - 2.342)^2 \\
 &= 12(2.840 - 2.342)^2 + 12(1.744 - 2.342)^2 = 8.5753,
 \end{aligned}$$

$$\begin{aligned}
 SS_E &= SS_{\text{total}} - SS_{\text{subject}} - SS_{\text{phase}} - SS_{\text{treatment}} \\
 &= 109.7462 - 60.5631 - 19.7871 - 8.5753 = 20.8207.
 \end{aligned}$$

Table 22.18 ANOVA table for losing weights.

Source	DF	SS	MS	F	P
Among objects	11	60.5631	5.5057	2.64	> 0.05
Between periods	1	19.7871	19.7871	9.50	< 0.05
Between treatments	1	8.5753	8.5753	4.12	> 0.05
Error	10	20.8207	2.0821		
Total	23	109.7462			

Program 22.1 Program for analyzing Example 22.1.

Line	Program	Line	Program
01	DATA FE;	09	PROC MEANS;
02	INPUT A B X @@;	10	VAR X;
03	CARDS;	11	CLASS A B;
04	1 1 10 1 1 10 1 1 40 1 1 50 1 1 10	12	RUN;
05	1 2 30 1 2 30 1 2 70 1 2 60 1 2 30	13	PROC ANOVA;
06	2 1 10 2 1 20 2 1 30 2 1 50 2 1 30	14	CLASS A B;
07	2 2 50 2 2 50 2 2 70 2 2 60 2 2 30	15	MODEL X = A B A*B;
08	;	16	RUN;

Using Table 22.16 (or statistical software) to create ANOVA table for two-phase cross-over design (Table 22.18).

From the P -values in Table 22.18, it cannot be concluded that the difference between the two drugs a_1 and a_2 is statistically significant, but the losing weights at phase I is much more than those at phase II ($P < 0.05$).

This section only concerns with the comparison between two treatments. The comparison among three treatments or more than three treatments needs to use multi-phase cross-over design, for which readers with interest can refer to other textbooks or references.

22.4 Computerized Experiments

Experiment 22.1 Two-way analysis of variance for 2×2 factorial experiment with complete randomized design (see Program 22.1).

Experiment 22.2 Analysis of variance for completely randomized split-plot design In Program 22.2, R , A , B and MM represent numbers of replications, levels of the factor for whole-plots, levels of the factor for

Program 22.2 Program for analyzing Example 22.2.

Line	Program	Line	Program
01	DATA SPLIT;	13	5 2 1 20.75 5 2 2 23.75
02	INPUT R A B MM @@;	14	;
03	CARDS;	15	PROC MEANS;
04	1 1 1 15.75 1 1 2 19.00	16	VAR MM;
05	2 1 1 15.50 2 1 2 20.75	17	CLASS A B;
06	3 1 1 15.50 3 1 2 18.50	18	RUN;
07	4 1 1 17.00 4 1 2 20.50	19	PROC ANOVA;
08	5 1 1 16.50 5 1 2 20.00	20	CLASS R A B;
09	1 2 1 18.25 1 2 2 22.25	21	MODEL MM = A R*A B A*B;
10	2 2 1 18.50 2 2 2 21.50	22	TEST H = A E = R*A;
11	3 2 1 19.75 3 2 2 23.50	23	RUN;
12	4 2 1 21.50 4 2 2 24.75		

Program 22.3 Program for analyzing Example 22.4.

Line	Program	Line	Program
01	DATA CROSS;	13	10 1 2 0.227 10 2 1 1.271
02	INPUT P TIME TREAT WEIGHT @@;	14	11 1 2 1.589 11 2 1 1.271
03	CARDS;	15	12 1 2 0.136 12 2 1 1.589
04	1 1 1 6.129 1 2 2 -0.454	16	;
05	2 1 1 2.497 2 2 2 0.908	17	PROC MEANS;
06	3 1 1 4.313 3 2 2 0.454	18	CLASS TIME TREAT;
07	4 1 1 4.540 4 2 2 2.724	19	VAR WEIGHT;
08	5 1 1 1.498 5 2 2 1.135	20	RUN;
09	6 1 1 8.172 6 2 2 4.313	21	PROC ANOVA;
10	7 1 2 4.449 7 2 1 2.043	22	CLASS P TIME TREAT;
11	8 1 2 4.994 8 2 1 1.816	23	MODEL WEIGHT = P TIME TREAT;
12	9 1 2 0.454 9 2 1 0.136	24	RUN;

subplots and the observations of the treatments respectively. Note that factor *A* and whole-plots error must be indicated in line 22. Program 22.2 can be modified for randomized complete-block split-plot design by changing line 21 as MODEL MM = R A R*A B A*B.

Experiment 22.3 Analysis of variance for two-phase cross-over design (see Program 22.3).

22.5 Practice and Experiments

1. In medical study, the experiment factor usually has two options, such as “operation” and “no operation”, “chemotherapy” and “no chemotherapy”. Try to design a 2^2 factorial experiment and explain the positive interaction (cooperation effect) and negative interaction (opposition effect).
2. How to use completely random design, randomized complete-block design and Latin-square design to arrange factorial treatment.
3. Take 2^2 factorial designs as examples to figure out the head of analysis of variance table (source and DF) for completely random design, randomized complete-block design and Latin-square design.
4. Perform residual analysis after square root or logarithm transformation for Example 22.1, and check if the independence and normality have improved.
5. State the relation and difference between factorial design and split-plot design.
6. Study the breath resistance in different loads (factor A) and different aviation oxygen supply device (factor B). Factor A has three levels, sitting quietly, 250 and 600 kg/min body force load; factor B has two levels, YX-1 and YX-2 oxygen supply systems. There are 12 objects, each of which can only receive one treatment. How to design the experiment? If in a period of time, each person can receive repeatedly all the treatments, how to design the experiment? Try to write the randomized group result.
7. Compare the repair effect of three constitutional drugs (factor A) for the neural lesion and two neural suture methods, the experiment objects are 12 dogs; two persons can help to measure the neural lesion. How to design the experiment? And write down a randomized allocation table.
8. It is the result of completely randomized factorial design in Table 22.19, in which factor A is the exposure frequency of millimeter waves, and factor B is the exposure time, and T_k ($k = 1, 2, \dots, 15$) is the sum of each treatment group. Try to work out the statistical analysis.
9. Ten moderate hyperthyroid patients were randomly divided into two groups, and methidathion and methidathion plus inderal were used to cure them respectively. The measuring results of heart rate before therapy and four weeks after therapy were given in Table 22.20.

Table 22.19 Contents of DNA in liver cells of mice ($r = 5$) (AU , absolutely unit).

Factor $B(J = 5)$	Factor $A(I = 3)$			Factor B total (B_i)
	36.04 GHz	50.05 GHz	Control	
Instant	2.203 5 (T_1)	1.938 0 (T_6)	2.182 0 (T_{11})	6.323 5 (B_1)
1 d	1.915 5 (T_2)	1.914 0 (T_7)	1.987 5 (T_{12})	5.817 0 (B_2)
3 d	1.970 5 (T_3)	1.663 0 (T_8)	1.882 5 (T_{13})	5.516 0 (B_3)
5 d	1.912 0 (T_4)	1.981 0 (T_9)	2.061 5 (T_{14})	5.954 5 (B_4)
7 d	1.924 0 (T_5)	1.975 5 (T_{10})	1.909 0 (T_{15})	5.808 5 (B_{15})
Factor A total (A_i)	9.925 5 (A_1)	9.471 5 (A_2)	10.022 5 (A_3)	29.419 5 ($\sum X$)
Sum of squares	$\sum X^2 = 11.7074$			

From: XiaoJuan Wang. Study about the effect of millimeter waves to liver of little rat. Disi Junyi Daxue Xuebao (J of Fourth Military Med University), 1990, 11(2): 92.

Table 22.20 Heart rate before and after therapy of hyperthyroid (time/min).

Curing methods	Before therapy	Four weeks after therapy
Methidathion	115	91
	120	94
	124	88
	116	82
	114	96
Methidathion + inderal	117	83
	110	80
	118	92
	119	85
	122	84

- Disassemble variances and analyze main effect of treatment method and time, and their interaction.
- 10. To analyze the block SS of the first part of whole-plots units of split-plot design in Table 22.3, and provide table of analysis of variance for the completely randomized split-plot design.
 - 11. To compare the measurement results of oxygen consumption between two devices A and B (treatment factor), 14 healthy persons with similar condition were tested. The objects and sequence of test were two important none-treatment factors. Each person was tested by the two

Table 22.21 Results of oxygen consumption between devices *A* and *B* (ml/h).

Objects number	Phase I		Phase II	
	Device	Oxygen consumption	Device	Oxygen consumption
1	<i>A</i>	1 237	<i>B</i>	1256
2	<i>B</i>	1 387	<i>A</i>	1348
3	<i>A</i>	1179	<i>B</i>	1275
4	<i>B</i>	1025	<i>A</i>	1022
5	<i>B</i>	1225	<i>A</i>	1226
6	<i>A</i>	1000	<i>B</i>	981
7	<i>B</i>	1050	<i>A</i>	1026
8	<i>A</i>	1295	<i>B</i>	1387
9	<i>A</i>	1218	<i>B</i>	1187
10	<i>B</i>	1050	<i>A</i>	1031
11	<i>A</i>	1138	<i>B</i>	1175
12	<i>B</i>	1387	<i>A</i>	1298
13	<i>B</i>	1150	<i>A</i>	1108
14	<i>A</i>	971	<i>B</i>	1012

From: Shuqin Yang, Zuchao Guo. China Medicine Encyclopedia (medical statistics), 1985.

devices in order. The 14 persons were paired according to their similarity of conditions; randomly selected one person of the pair to use devices *A* first and then *B*, and the counterpart to use in the other way. Work out a statistical analysis for the data in Table 22.21.

(1st edn. and 2nd edn. Yongyong Xu, Yi Wan, Jiqian Fang)



Chapter 23

Analysis of Repeated Continuous-Type Measurements

Repeated measure data refer to multiple measurements of the same variable taken from the same experimental unit or subject (human being, animal, equipment, etc.) at different times. The response variable from repeated measures may be continuous, discrete or binary. Analysis of discrete and binary data requires advanced statistical methodology like generalized estimating equations, GEEs. Interested readers can refer to other referential materials. In practice, continuous type repeated measurements are common. In this chapter we will discuss some statistical methods for analyzing continuous type repeated measure data.

23.1 Examples of Repeated Measurements

The experiment with repeated measurements usually concerns both treatment and time factors. The following lines should be directed towards the design: subjects are assigned to different treatment groups at random, and it is better to use parallel control group; the time points are designated in advance; the measurements at time 0, i.e., the time point just before the treatment be administered, is used as baseline; and each subject receives measures regularly.

Example 23.1 A nutrition experiment was conducted to explore the effect of test food on serum cholesterol. Seven rabbits in each of the two groups were fed with normal-food and test-food, respectively. The serum cholesterol concentration (mmol/L) was measured just before and after five,

Table 23.1 Logarithmic-transformed serum cholesterol concentration (mmol/L) from a nutrition experiment.

Treatment group (Test food)				Control group (Normal food)			
Rabbit	Before experiment	5 weeks after	10 weeks after	Rabbit	Before experiment	5 weeks after	10 weeks after
1	0.744741	2.013341	2.621341	8	0.375741	0.667841	0.569941
2	0.904141	2.054141	1.628441	9	0.994741	0.584441	0.461241
3	0.357641	1.137841	2.196741	10	0.598841	0.955541	0.598841
4	1.077741	1.948741	2.239241	11	0.719741	1.354241	1.032441
5	0.584441	1.668441	0.985041	12	0.157041	0.246141	0.613041
6	0.985041	1.926241	2.915641	13	0.861241	0.882941	0.757041
7	1.050841	1.638641	1.225541	14	0.872141	0.555041	0.540041

From: Xu Y. (1987). Journal of Chinese Preventive Medicine. 21(1):34.

Table 23.2 Blood concentration ($\mu\text{mol/L}$) of two drug forms at four time points.

Old form					New form				
Subject	0 hour	4 hours	8 hours	12 hours	Subject	0 hour	4 hours	8 hours	12 hours
1	90.53	142.12	65.54	73.28	8	70.53	97.38	112.12	58.50
2	88.43	163.17	48.95	71.77	9	68.43	95.27	133.17	56.90
3	100.01	144.75	86.06	80.01	10	57.37	78.43	83.16	48.34
4	46.32	126.33	48.95	39.54	11	105.80	120.54	136.33	84.03
5	73.69	138.96	70.02	60.89	12	80.01	104.75	114.75	65.61
6	105.27	126.33	75.01	83.66	13	56.32	75.27	96.33	47.52
7	86.32	121.06	78.95	70.24	14	53.69	110.02	138.96	45.44
					15	85.27	110.01	126.33	69.47
					16	66.32	115.27	129.06	55.29

ten weeks of experiment. The logarithmic-transformed data are listed in Table 23.1.

Example 23.2 A pharmaceutical study was conducted to explore the metabolic difference between two dosage forms, old and new, of a drug. Blood concentration of the drug were measured at 0, 4, 8 and 12 hours after administration on 16 subjects, 7 for old form and 9 for new. Data are listed in Table 23.2. The question is whether there is a significant difference between the dose-time curves.

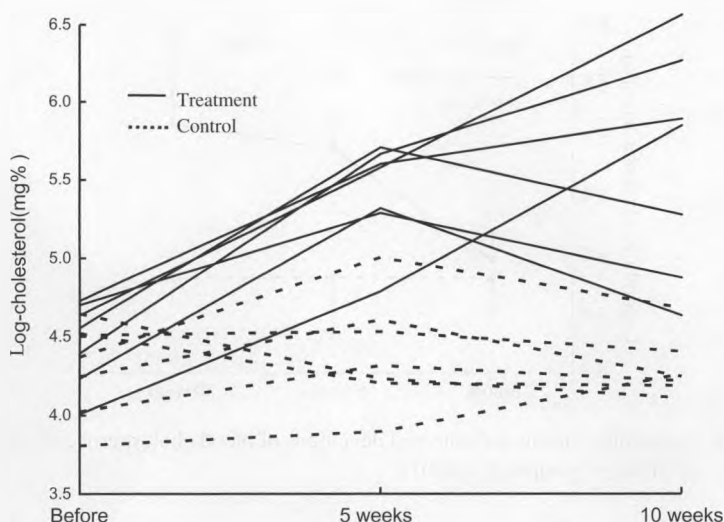


Fig. 23.1 Blood cholesterol changes of rabbits with time in two groups.

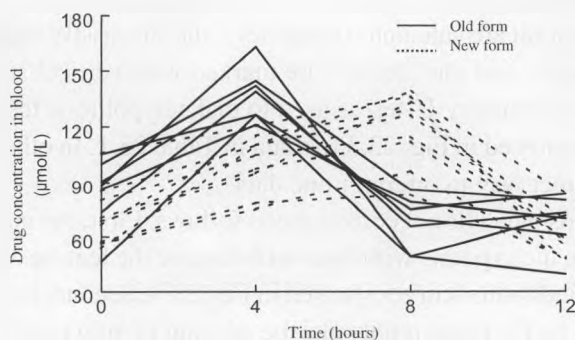


Fig. 23.2 Drug concentration changes with two for two forms.

From one data in Tables 23.1 and 23.2, Figs. 23.1 and 23.2 have been drawn. In these graphs the horizontal axis represents the time scale and the vertical axis represents the response level. Each curvilinear represents a time trend of responses for one subject. Different types of lines are used to discriminate different groups. This graph can give a rough imagination of the differences between groups and the changes of response with time individually.

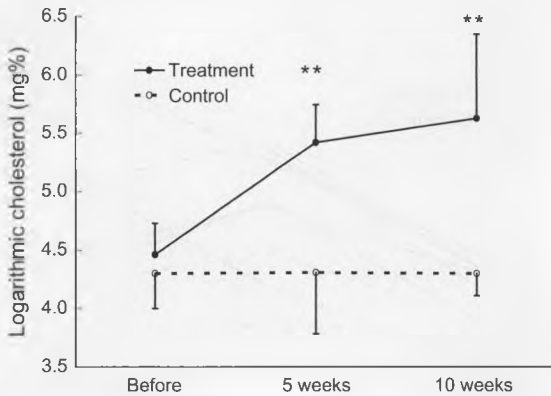


Fig. 23.3 Logarithmic means and standard deviations of blood cholesterol concentrations in rabbits for different groups ($P < 0.01$).

23.2 Imperfect Analysis and its Origins

A common mistake to handle repeated measure data is to treat them as independent observations: the mean and standard deviation of observations for each time point are calculated separately; the successive mean points are linked with lines, and the “errors” are marked with vertical sticks; and the t -test or Mann–Whitney U -test is used to make hypothesis testing for each time point as showed in Fig. 23.3 for data in Table 23.1. In other words, this is not the correct way to interpret one data.

The reasons why it is not correct are as follows. First, the method linking the successive mean points with lines will obscure the feature of the location and shape of individual curves showed in Fig. 23.1. Second, the shape of the curve formed by the mean points may be relevant to individual’s curvilinear shapes. Third, Fig. 23.3 depicts the corresponding standard deviation for each time point to show that the two bunches of curves are located closely around their means respectively and to reflect that the two bunches of curves are separated greatly because the two sticks representing standard deviations do not overlap. But it is not true. Finally, the method mentioned above does not reflect the fact that the measurements at different time points come from the same individual and are correlated to each other (Tables 23.3 and 23.4).

Table 23.4 shows that the correlation coefficient is 0.507 between the measures before and five weeks after experiment. And the value is 0.777 between the measures five and ten weeks after experiment.

Table 23.3 Variance-covariance matrix of measurements between time points (from data in Table 23.1).

	Before	5 weeks	10 weeks
Before	0.081	0.090	0.065
5 weeks		0.386	0.411
10 weeks			0.723

Table 23.4 Correlation coefficients of measurements between time points (from data in Table 23.1).

	Before	5 weeks	10 weeks
Before	1	0.507	0.269
5 weeks		1	0.777
10 weeks			1

In fact, the final reason is the key for repeated data analysis. Repeated measures in each subject correspond to a matched experiment with oneself as control. These repeated measurements for the same subject have fine comparability. Due to the correlation between repeated measurements collected from the same subject, one should account for this when analyzing. Otherwise the statistical conclusion may be doubtful. In other words, the core to analyze repeated measurements is how to handle such feature of correlation ingeniously.

23.3 Approach with Summary Measures

To avoid the correlation among repeated measurements, a new set of fewer independent measures can be taken to summarize as much as possible information existing in the correlated original repeated measurements. Then the univariate statistical methods like *t*-test, ANOVA or Mann-Whitney *U*-test can be used directly to the new set of independent measures for comparison among groups. The methodology is known as the summary measures approach or derived variable approach. Through the summary measures approach is easy to carry out, the selection of summary measures is not easy. So the summary measures should have clear-cut clinical or biological significance and should be specified in design stage.

Table 23.5 The common used summary measures of repeated measurements for choice.

Data type	Topic of interest	Summary measures
Peak	Are there differences between treatment groups?	Means (equal time intervals); Areas under curves (Unequal time intervals)
Peak	Is there difference in maximum (or minimum) responses between treatment groups?	Maximum (or minimum) value
Peak	Is there difference between times achieving maximum (or minimum) responses?	Time to achieving maximum (or minimum) responses
Growth	Are changing rates in different treatment groups equal?	Regression coefficients
Growth	Are final results of treatment groups equal?	End value of response; difference of end value from start value; changing rate between end value and start value.
Growth	Are response rates of different treatment groups equal?	Waiting time arriving at a specific value (for example, time from baseline to a specified times of baseline value)

The data of repeated measurements in practice mostly include two types, peak (peak value) type and growth (monotone increase or decrease) type according to the trend of repeated measurements with time. Based on the trend and topic of interest, the researchers often select the summary measures as listed in Table 23.5.

From the data in Table 23.1 every subject has a mean which is the average of original measurements over three time points. For example, the mean of rabbit 1 is $(0.7447 + 2.0133 + 2.6213)/3 = 1.7931$. By using *t*-test to compare the two sets of summary measures, the results show that the means of treatment and control groups are significantly different ($p < 0.001$). The results are listed in Table 23.6.

23.4 Analysis of Variance for Repeated Measurements

The method introduced to deal with the data in Table 23.1 is a common technique, where the mean of measurements on each subject is used as a characteristic for comparison between groups. In general, if we only want to compare the differences between two or more group means, the mean

Table 23.6 Summary measures and results of *t*-test for data in Table 23.1.

Group	Summary measure (Mean for each rabbit)				Grand mean	<i>t</i> -value	<i>df</i>	<i>P</i> -value
Treatment	1.79314	1.52891	1.23074	1.75524	1.5192	5.6295	12	0.0001
	1.07931	1.94231	1.30501					
Control	0.53784	0.68014	0.71774	1.03547	0.6856			
	0.33874	0.83374	0.65574					

or sum of repeated measurements may be used as a summary measure for univariate hypothesis testing. But in doing so, the possible differences in time trends between groups cannot be revealed. In view of this, we introduce the analysis of variance for repeated measurements here.

Data in Table 23.1 look like a randomized block design. But the difference from random block design is that the measurements are arranged with time order, not arranged randomly as in the random block design, and the readings at different time points are correlated to each other in some degree. Generally, the degree of correlation varies with the distance between the two measured time points. It is necessary to take action according to the correlation type.

A statistical model for the data as showed in Table 23.1 is

$$Y_{git} = \mu + \alpha_g + \delta_{i(g)} + \beta_t + (\alpha\beta)_{gt} + \epsilon_{git}, \quad (23.1)$$

Y_{git} is the measurement of response variable at time t ($t = 1, \dots, q$) on the i th ($i = 1, \dots, n_g$) subject in the group g ($g = 1, \dots, m$); n_g is the total number of group g ; μ is the expectation of its population; α_g is the treatment effect for group g ; $\delta_{i(g)}$ is the effect associated with subject i in group g ; β_t is the effect at time t . $(\alpha\beta)_{gt}$ is the interaction of GROUP*TIME for group g with time t ; ϵ_{git} is the random error associated with the i th subject assigned to group g at time t .

If the estimate of the parameter $(\alpha\beta)_{gt}$ is statistically significant, it reflects the fact that the response strength varies with group and time point. In other words, the analysis of variance model for repeated measurement data is the sum of effect components resulted from different sources, especially for time effects which include the main effect of time β_t and the interaction $(\alpha\beta)_{gt}$.

However, it can be proved that if the correlation matrix meets certain special conditions, the idea of ANOVA for random block design can still be applied to the analysis of repeated measurements data. The commonly used precondition is the so-called "sphericity". For example, suppose there are t observed time points, if the correlations between the readings at time points 1 and 2, 2 and 3, \dots , $t - 1$ and t , t and 1, \dots , are kept the same, then the property of "sphericity" holds.

23.4.1 Test for sphericity of variance-covariance matrix

The variance-covariance and correlation matrices for data in Table 23.1 are listed in Tables 23.3 and 23.4 respectively. We use statistical software to test the "sphericity" of the variance-covariance matrix. As there may exist auto-correlation between measurements within the same subject, i.e., dependence among measurements. This can affect the reliability of statistical inference. Let Σ represents the variance-covariance matrix, I represents a unity matrix, the hypothesis is expressed as:

$$H_0 : \Sigma = \sigma^2 I, \quad H_1 : \Sigma \neq \sigma^2 I.$$

Mauchly's test is used for "sphericity". If the P -value of the test is large (for example, greater than 0.05), the null hypothesis of sphericity will not be rejected. In this situation we can deal with the time as a single variable, each time point as a level of the time variable. The uni-variate t -test or variance analysis may be valid. When the sphericity assumption is false, it is still possible to use the ANOVA but with adjusted P -values by G-G (Greenhouse-Geisser) procedure or H-F (Huynh-Feltd) procedure.

The test statistic for sphericity of data in Table 23.1 is $\chi^2 = 5.1628$ approximately with 2 degrees of freedom and the corresponding P -value is 0.0757, leading to that the "sphericity" hypothesis would not be rejected. Thus time can be treated as a single variable and a unit-variate ANOVA can be applied.

23.4.2 Univariate analysis of variance

Suppose that we have m treatment groups and q time points. n_g is the sample size of group g ($g = 1, 2, \dots, m$), the total subjects

$$n = \sum_{g=1}^m n_g.$$

Let Y_{git} be an observation on subject i ($i = 1, 2, \dots, n_g$) at time point t ($t = 1, \dots, q$) for group g . We can obtain the following averages:

Grand mean:

$$\bar{y} = \left(\sum_{g=1}^m \sum_{i=1}^{n_g} \sum_{t=1}^q y_{git} \right) / n. \quad (23.2)$$

Treatment mean for group g :

$$\bar{y}_g = \left(\sum_{i=1}^{n_g} \sum_{t=1}^q y_{git} \right) / (q \times n_g), \quad g = 1, \dots, m. \quad (23.3)$$

Mean of individual i in group g :

$$\bar{y}_{i(g)} = \left(\sum_{t=1}^q y_{git} \right) / q, \quad i = 1, \dots, n_g; \quad g = 1, \dots, m. \quad (23.4)$$

Mean at time t :

$$\bar{y}_t = \left(\sum_{g=1}^m \sum_{i=1}^{n_g} y_{git} \right) / \left(\sum_{g=1}^m n_g \right), \quad t = 1, \dots, q. \quad (23.5)$$

Mean at time t for group g :

$$\bar{y}_{gt} = \left(\sum_{i=1}^{n_g} y_{git} \right) / n_g, \quad t = 1, \dots, q; \quad g = 1, \dots, m. \quad (23.6)$$

From these averages calculated above, the parameters in the model can be estimated. For example:

$$\begin{aligned} E(\bar{y}_g - \bar{y}) &= \alpha_g, & E(\bar{y}_{i(g)} - \bar{y}_g) &= \delta_{i(g)}, & E(\bar{y}_t - \bar{y}) &= \beta_t, \\ E(\bar{y}_{gt} - \bar{y}_g - \bar{y}_t + \bar{y}) &= (\alpha\beta)_{gt}, & E(\varepsilon_{git}) &= \bar{y}_{git} - \bar{y}_{i(g)} - \bar{y}_{gt} + \bar{y}_g. \end{aligned}$$

According to the principle of analysis of variance, under the condition that

$$\sum_{g=1}^m \alpha_g = \sum_{i=1}^{n_g} \delta_{i(g)} = \sum_{t=1}^q \beta_t = \sum_{g=1}^m (\alpha\beta)_{gt} = \sum_{k=1}^q (\alpha\beta)_{gt} = 0$$

the partitioning of the total variation is accomplished by separating the total sum of squares SS_T for the observed data. The total degrees of freedom can

also be partitioned according to its source. The mean square (MS) values are the sum of squares divided by the corresponding degrees of freedom. Various sum of squares of the differences from its mean and ANOVA Table are showed in Table 23.7.

The SS in column 3 of Table 23.7 is acronym referring to the sum of squared differences from its mean.

Different from the general ANOVA, which has only one error term (SS_{subj}), here Table 23.7 for repeated measurements gives two error terms, error between subjects and error within subject. The latter is the error corresponding to time and the interaction between treatment and time, denoted with “treatment \times time” for short.

Example 23.3 By using formulas showed in Table 23.7, the data in Table 23.1 are dealt with univariate analysis of variance.

Solution We have $n = 14$, $m = 2$, $q = 3$, $L = mq = 6$ and $n_g = n_1 = n_2 = r = 7$. U_i , G_g , T_t , and GT_l are the sums of measurements for individual i , for group g , for time point t and for subgroup l respectively (see Table 23.8). By computer software, we have the results showed in Table 23.9.

From Table 23.9 the F -statistic for differences between treatment and control groups is 31.69 with degrees of freedom 1 and 12, corresponding to a P -value of 0.0001; the F -statistics for time is 11.93 with degrees of freedom 2 and 41, corresponding to a P -value of 0.0003; and the F -statistics for the interaction of group \times time is 10.57 with degrees of freedom 2 and 41, corresponding a P -value of 0.0005. The results show a significant interaction between food and time, that is, the food of interest has significant effect on the time trend of blood cholesterol concentration.

23.4.3 Analysis of variance for data with orthogonal transformation

The software SAS usually gives two results for sphericity test. Sometimes the two results may be different. If the first result gives a P -value greater than 0.05, the univariate analysis of variance can be applied for the original

Table 23.7 Table of uni-variate analysis of variance.

Source	DF	SS	MS	F
Between treatments	$m - 1$	$SS_{\text{trt}} = \sum_{g=1}^m q n_g (\bar{y}_g - \bar{y})^2$	$MS_{\text{trt}} = \frac{SS_{\text{trt}}}{m - 1}$	$F_{\text{trt}} = \frac{MS_{\text{trt}}}{MS_{\text{subj}}}$
Between subjects	$n - m$	$SS_{\text{subj}} = \sum_{g=1}^m \sum_{i=1}^{n_g} q (\bar{y}_{i(g)} - \bar{y}_g)^2$	$MS_{\text{subj}} = \frac{SS_{\text{subj}}}{n - m}$	
Time	$q - 1$	$SS_t = \sum_{i=1}^q \sum_{g=1}^m n_g (\bar{y}_t - \bar{y})^2$	$MS_t = \frac{SS_t}{q - 1}$	$F_t = \frac{MS_t}{MS_{\text{within}}}$
Treatment \times Time	$(m - 1)(q - 1)$	$SS_{\text{trt} \times t} = \sum_{g=1}^m \sum_{i=1}^q n_g (\bar{y}_{gt} - \bar{y}_g - \bar{y}_t + \bar{y})^2$	$MS_{\text{trt} \times t} = \frac{SS_{\text{trt} \times t}}{(m - 1)(q - 1)}$	$F_{\text{trt} \times t} = \frac{MS_{\text{trt} \times t}}{MS_{\text{within}}}$
Within subject	$(n - m)(q - 1)$	$SS_{\text{within}} = \sum_{g=1}^m \sum_{i=1}^{n_g} \sum_{t=1}^q (y_{git} - \bar{y}_{i(g)} - \bar{y}_{gt} + \bar{y})^2$	$MS_{\text{within}} = \frac{SS_{\text{within}}}{(n - m)(q - 1)}$	
Total	$nq - 1$	$SS_{\text{tot}} = \sum_{g=1}^m \sum_{i=1}^{n_g} \sum_{t=1}^q (y_{git} - \bar{y})^2$		

Table 23.8 Natural logarithmic transformation of blood concentration (mg%) of the rabbits.

Rabbit					Rabbit				
Treatment					Control				
(i)	Before	5 weeks	10 weeks	U_i	(i)	Before	5 weeks	10 weeks	U_i
1	0.744741	2.013341	2.621341	5.379424	8	0.375741	0.667841	0.569941	1.613524
2	0.904141	2.054141	1.628441	4.586724	9	0.994741	0.584441	0.461241	2.040424
3	0.357641	1.137841	2.196741	3.692224	10	0.598841	0.955541	0.598841	2.153224
4	1.077741	1.948741	2.239241	5.265724	11	0.719741	1.354241	1.032441	3.106424
5	0.584441	1.668441	0.985041	3.237924	12	0.157041	0.246141	0.613041	1.016224
6	0.985041	1.926241	2.915641	5.826924	13	0.861241	0.882941	0.757041	2.501224
7	1.050841	1.638641	1.225541	3.915024	14	0.872141	0.555041	0.540041	1.967224
GT_i	5.704589	12.38739	13.81199	31.90397		4.579489	5.246189	4.572589	14.39827
$G_1 = 5.704589 + 12.38739 + 13.81199 = 31.90397$									
$G_2 = 4.579489 + 5.246189 + 4.572589 = 14.39827$									
$T_1 = 5.70489 + 4.579489 = 10.28408 \quad T_2 = 12.38739 + 5.246189 = 17.63358$									
$T_3 = 13.81199 + 4.572589 = 31.90397$									

Table 23.9 Univariate analysis of variance table.

Source	DF	SS	MS	F	P
Group	1	7.2964	7.2964	31.69	0.0001
Error between subjects	12	2.7628	0.2302		
Time	2	2.8618	1.4309	11.93	0.0003
Group \times time	2	2.5342	1.2671	10.57	0.0005
Error within subjects	24	2.8781	0.1199		
Total	41	18.3333			

observations. If the first result gives a P -value less than 0.05 and the second result gives a P -value greater than 0.05, then some orthogonal transformation for the original observations is needed before carrying out the univariate analysis of variance.

There are four kinds of commonly used orthogonal transformations:

1. Polynomial transformation. It is often used when the intervals of time points are not equal, such as weeks 1, 2, 5, 10 at which repeated measures are done. This transformation can reflect whether there is a linear, quadratic, or cubic trend of measurements with time.

Table 23.10 Results of testing trend with polynomial contrast.

Source	Contrast variable: WEEK_1(Linear)				Contrast Variable: WEEK_2(Quadratic)			
	DF	SS	F	P	DF	SS	F	P
Mean	1	2.3435	13.87	0.0029	1	0.5183	7.31	0.0192
Group	1	2.3515	13.92	0.0029	1	0.1827	2.58	0.1343
Error	12	2.0275			12	0.8506		

2. Helmert transformation. It is used to compare the mean of a measure at a time point to the means of subsequent measures. It is useful when determining a stable time point.
3. Contrast or simple transformation. A time point is selected in advance as a control level, against which the others are compared. Usually, researches take the baseline as a control level.
4. Profile transformation. It is used for comparison between two adjoining time points. This transformation is often used when polynomial transformation is unreasonable.

Now we use a polynomial transformation for the data in Table 23.1 to test whether there appears a linear or quadratic trend with time and whether there is any difference in time trend between the two groups. The number of time points are 3 thus we can fit up to a second-degree ($3-1 = 2$) polynomial time trends of serum cholesterol concentration for each of the two groups. The results of the test are shown in Table 23.10. In the table, the "Contrast Variable: WEEK_1" represents linear trend, and the "Contrast Variable: WEEK_2" quadratic trend. The test for "Mean" is referred to test if the means under the trend are equal to zero, which is equivalent to evaluation of the effect of time. It is showed that the "Mean" for both linear and quadratic contrasts are statistically significant ($p = 0.0029$ and 0.0192 respectively). The test for "Group" is really to test the interaction between group and time. It is showed that the linear contrast is significant ($p = 0.0029$) but quadratic contrast is not ($p = 0.1343$). These testing results indicate that there is a significant difference in linear trend, but none in quadratic contrast between the two groups.

Table 23.11 shows the means in different time points for treatment group and control group. We can see that there is an ascending trend for the treatment group, but no such trend for the control group.

Table 23.11 Means at different time points for two groups.

Group	Before experiment	5 weeks	10 weeks
Treatment	0.8149	1.7696	1.9731
Control	0.6542	0.7495	0.6532

23.4.4 Multivariate analysis of variance

The multivariate analysis of variance can be used to handle the repeated measures data regardless of sphericity condition. The method deals with each time point as an independent variable, and deals with a set of measurements over time points for one subject as a vector. Four test statistics are used for multivariate hypothesis testing. They are Wilks' Lambda, Pillai's Trace, Hotelling–Lawley Trace, and Roy's Greatest Root (see Chap. 14). Table 23.12 shows the results of multivariate analysis of variance for the data in Table 23.1, which are summarized based on the output of the SAS program.

For the effect of time, four test statistics are equivalent in $F = 19.0622$ with degrees of freedom 2 and 11, with a P -value of 0.003. Consequently, in view of the small value of P , we conclude a significant change of blood cholesterol concentrations with time for both treatment and control groups. The very small P -value of 0.0010 for interaction between time and group leads to a conclusion that the time trend of blood cholesterol concentration in the treatment group is significantly different from that in the control group.

As showed above, the three different methods (summary measures, univariate analysis of variance, and multi-variate analysis of variance) are applied to the data in Table 23.1, and reach the same conclusion. In general, the approaches of summary measures and multi-variate analysis of variance can be applied to any continuous repeated measurements. But when sphericity condition is satisfied (P -value for sphericity testing greater than 0.05), it is advisable to adopt univariate analysis of variance.

In fact, not all data can be dealt with univariate analysis of variance. For data in Example 23.2 (Table 23.2), the approximate χ^2 statistic of sphericity testing is 27.0284 with degrees of freedom 5, leading to a P -value of 0.0001. The results show that the univariate analysis of variance is inferior to multi-variate analysis of variance for dealing with such a dataset.

Table 23.12 Multivariate analysis of variance for data in Table 23.1.

Effect	Statistic	Value	<i>F</i>	Num DF	Den DF	<i>Pr</i> > <i>F</i>
Time	Wilks' Lambda	0.2239	19.0622	2	11	0.0003
	Pillai's Trace	0.7761	19.0622	2	11	0.0003
	Hotelling-Lawley Trace	3.4658	19.0622	2	11	0.0003
	Roy's Greatest Root	3.4658	19.0622	2	11	0.0003
Time × Group	Wilks' Lambda	0.2862	13.7157	2	11	0.0010
	Pillai's Trace	0.7138	13.7157	2	11	0.0010
	Hotelling-Lawley Trace	2.4938	13.7157	2	11	0.0010
	Roy's Greatest Root	2.4938	13.7157	2	11	0.0010

23.5 Computerized Experiments

Experiment 23.1 The *t* test for means as summary measures In program 23.1, lines 01–19 read the data into SAS dataset REP; among them line 03 is to compute the mean of repeated measurements for each subject and is named YBAR. Lines 05–18 are original data in Table 23.1. Lines 20–22 invoke TTEST procedure to perform a *t*-test for variable YBAR. If more than two treatment groups are to be compared, the procedure GLM can be invoked to perform an analysis of variance. the procedure NPAR1WAY can also be used to perform a nonparametric test. The output from program 23.1 is listed in Table 23.6.

Program 23.1 Analysis of summary measures for data in Table 23.1.

Line	Program	Line	Program
01	DATA REP;	13	2 0.994741 0.584441 0.461241
02	INPUT GROUP Y1-Y3;	14	2 0.598841 0.955541 0.598841
03	YBAR = MEAN(OF Y1-Y3);	15	2 0.719741 1.354241 1.032441
04	CARDS;	16	2 0.157041 0.246141 0.613041
05	1 0.744741 2.013341 2.621341	17	2 0.861241 0.882941 0.757041
06	1 0.904141 2.054141 1.628441	18	2 0.872141 0.555041 0.540041
07	1 0.357641 1.137841 2.196741	19	;
08	1 1.077741 1.948741 2.239241	20	PROC TTEST;
09	1 0.584441 1.668441 0.985041	21	CLASS GROUP;
10	1 0.985041 1.926241 2.915641	22	VAR YBAR;
11	1 1.050841 1.638641 1.225541	23	PROC PRINT;RUN;
12	2 0.375741 0.667841 0.569941		

Program 23.2 General linear model for data of repeated measurements.

Line	Program
01	PROC GLM DATA = REP;
02	CLASS GROUP;
03	MODEL Y1-Y3 = GROUP/NOUNI;
04	REPEATED WEEK 3 (0 5 10) POLYNOMIAL
05	/SUMMARY PRINTE PRINTM;
06	LSMEANS GROUP/PDIFF;
07	RUN;

Experiment 23.2 General linear models for repeated measurements

Program 23.2 carries out the sphericity testing, univariate analyses of variance and multi-variate analyses of variance, which are introduced in Sec. 23.4 for the data in Table 23.1. Line 01 invokes the procedure GLM and specifies SAS dataset REP, which is created by program 23.1. Line 02 specifies GROUP as a categorical variable. Line 03 defines the model. Following MODEL statement, the response variables Y1–Y3, are put on the left of the equal sign, and independent variable GROUP (and other independent variables, if any) on the right of the equal sign. Option NOUNI suppresses the individual variance analysis of Y1, Y2 and Y3. In lines 04 and 05, the REPEATED statement handles the repeated measures design; WEEK is the name of a factor associated with the dependent variables; 3 is the number of levels of WEEK factor; the numbers in the parentheses are the values corresponding to the levels of WEEK factor. The option POLYNOMIAL defines the transformation. The default in SAS is the transformation CONTRAST; one can optionally specify an ordinal value of reference level in the parentheses for contrast following CONTRAST. The option SUMMARY prints the results of analysis of variance for transformed variables showed in Table 23.10, PRINTE asks sphericity test for transformed contrast, PRINTM asks to print the transformation matrix that defines the contrasts in the analysis, which is helpful in understanding the contrast. Line 06 asks to print the group means by time points (showed in Table 23.11). The sphericity test prints out two results, the former is for the original data, and the latter is for the orthogonal transformed components labeled “Applied to Orthogonal Components”.

Table 23.13 Symptom scores recorded on 20 patients at different time points.

Patient	Group	Before treatment	Days after treatment				
			10 days	60 days	120 days	180 days	270 days
1	1	0.60	0.67	2.84	2.10	2.00	1.60
2	1	1.42	3.40	4.10	2.92	2.65	3.40
3	1	0.90	2.30	2.70	1.70	1.10	1.30
4	1	1.10	1.40	1.00	2.60	0.90	2.10
5	1	2.30	2.20	3.80	3.50	2.50	1.80
6	1	0.81	1.20	1.12	1.61	1.49	1.61
7	2	1.20	1.10	1.13	3.49	1.57	1.54
8	2	2.71	2.04	2.61	2.17	2.15	1.81
9	2	1.02	1.43	1.61	1.70	2.82	1.55
10	2	1.71	1.71	1.21	0.90	0.61	1.66
11	2	1.16	0.78	0.51	0.85	0.88	0.49
12	2	0.85	1.25	1.66	2.13	1.04	0.62
13	2	0.60	2.50	2.20	1.20	1.11	1.00
14	2	0.90	0.80	0.70	1.00	0.80	0.60
15	2	3.40	3.30	3.40	3.40	2.10	1.50
16	2	1.10	1.20	1.50	2.40	1.50	3.20
17	2	4.60	1.20	3.20	2.30	2.30	1.50
18	2	1.60	0.90	1.80	2.10	1.30	1.10
19	2	0.40	0.96	1.01	0.71	0.59	0.60
20	2	1.80	1.40	1.00	1.30	2.40	2.40

From: Data adapted from Crowder, MJ and Hand, DJ. Analysis of repeated measures. (Chapman and Hall, 1990).

23.6 Practice and Experiments

1. In the analysis of repeated measurements by univariate ANOVA, what is the precondition when univariate hypothesis test is applied?
2. 20 patients were randomly assigned to one of the two treatment groups with six patients in group 1 and 14 patients in group 2. The patients in the same group received the same drug. Symptom scores were taken for each patient at the time points of before treatment and 10, 30, 120, 180, 270 days after treatment. The data are showed in Table 23.13.
 - (1) Draw a curvilinear graph with different types of lines representing the two groups. Time is used as horizontal axis and symptom scores as vertical axis. Observe whether there might be a statistically

- significant difference between the two groups, and whether a linear trend with time exists in the two groups.
- (2) By using the mean as a summary measure, test if the two groups are significantly different in their means.
 - (3) Analyze the data by univariate analysis of variance.
3. An experiment was conducted to study the effects of two nutritional elements on body weight. 16 mice were randomly divided into two groups. Mice in group *A* were fed with nutritional ingredient *A*, and those in group *B* with nutritional ingredient *B*. Their body weights (grams) were scaled at 0, 14, 28, 42, 49 and 63 days after the experiment had begun. The data are showed in Table 23.14. Compare the effects of the two ingredients for weight increments.
 4. Try to analyze data in Table 23.2 with single variate and multivariate ANOVA models. Find the difference between the results from the two methods. Decide which of the results should be adopted.

Table 23.14 Body weights (grams) at different days for 16 mice in two groups.

Group	Mouse No.	0 day	14 days	28 days	42 days	49 days	63 days
A	1	240	255	262	266	265	278
A	2	225	230	240	243	238	245
A	3	245	250	262	267	264	269
A	4	260	255	265	270	274	275
A	5	255	255	270	274	276	280
A	6	260	270	275	278	284	281
A	7	275	260	273	276	282	284
A	8	245	260	270	265	273	278
B	9	410	425	438	442	456	478
B	10	405	430	448	258	475	496
B	11	445	450	455	451	462	472
B	12	555	565	590	595	612	628
B	13	470	475	487	493	507	525
B	14	535	530	535	525	543	559
B	15	520	530	543	538	553	548
B	16	510	520	530	535	550	569

(1st edn. Songlin Yu; 2nd edn. Songlin Yu, Jiqian Fang)

Chapter 24

Design and Analysis of Cross-Sectional Studies

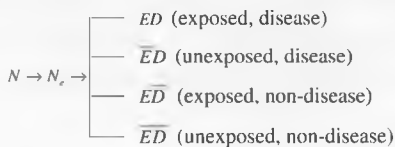
A cross-sectional study, also called a prevalence study or survey, adopts momentary observational methods to obtain information about the prevalence of disease infection and the exposure level of relevant factors in an effort to describe the relationship between a disease and important factors.

The goals of this type of studies are:

- (1) To find out how widely an existing disease prevails in different geographical areas, among different populations, and at different times, and what factors it is associated with, in order to determine who are at risk of contracting the disease. This provides evidence for further etiological studies.
- (2) To detect patients for early treatment and to prevent the disease from progressing.

24.1 Design of the Study

If both disease (D) and exposure (E) are dichotomous variables, a cross-sectional design model can be shown as the chart below:



where N denotes the population, N_e number of qualified individuals used in the study, E exposed to study factor, \bar{E} unexposed to study factor, D with interested disease, and \bar{D} without interested disease.

A cross-sectional study usually consists of the following components:

- (1) Purpose of the study. One should point out clearly the reason of conducting the study. The purpose is usually to explore the relationship between a disease and certain risk factors. For example, a study might be carried out “to discover the relationship between coronary heart disease and cholesterol level in blood serum”.
- (2) Subjects and sample size of a study. The subject refers to the target population being studied. Subjects are usually selected according to the goal and approach of the research. For example, in a study which tries to find out the relationship between coronary heart disease and the cholesterol level in blood serum, two different approaches may be adopted. One approach is to compare the cholesterol levels in patients with coronary heart disease with those in healthy people; another is to study the different incidences of the disease in people with different cholesterol levels.

The sample size estimation is discussed in Sec. 24.3. The standard diagnostic criteria of disease should be adopted before the study in order to avoid diagnostic error.

- (3) Variables to be observed. In practice, quantitative variables should be used as far as possible, because they are more accurate than qualitative ones.

If the variables are physiological or biochemical factors, they should be examined in the same way; if they are psychological or behavioral ones, standardized questionnaires or scales should be used. Diagnosis of disease should be based on well-defined criteria.

There are two main types of variables: disease-related variables and exposure or factors-related variables. In addition to these, other variables should also be set up to account for the confounding.

24.2 Sampling Methods and Estimation of Population Parameters

We will introduce some commonly used sampling methods. They are simply random sampling, systematic sampling, stratified sampling and cluster sampling. A population with finite number of individuals is called finite

population; in contrast, a population with infinite number of individuals is called infinite population. Formula of parameter estimation varies with sampling method. Under the same sampling frame the formula of parameter estimation differs for finite population from infinite population. A finite population, from which the sample size is much smaller than the total number of the population, is approximately treated as infinite population sometimes.

24.2.1 Simple random sampling

24.2.1.1 Sampling method

For example, if we want to obtain a random sample of 300 students from all the 3000 elementary school students in an area in order to study their prevalence of roundworm infection, we may assign a set of ID numbers to all the 3000 students; write each number on a piece of paper, and mix them in a box; then 300 numbers are drawn randomly such that those 300 students corresponding to these numbers form a random sample of our study.

We can also use a computer or calculator to generate random numbers. In the above example, we may instruct the computer to generate more than 300 4-digit numbers randomly such as 1716, 1818, 7650, 8619, Some of these numbers will be less than 3000, while others are larger than 3000. Those ranging from 1 to 3000 are kept unchanged; those above 3000 are changed into numbers that are less than or equal to 3000. This is done by subtracting 3000 from numbers ranging from 3000 to 6000, or subtracting 6000 from those ranging from 6000 to 9000. Those above 9000 are invalid numbers and discarded. We keep doing this until 300 different numbers are obtained such that the students corresponding to those numbers form our sample.

In simple random sampling, it is required for all the members of an interested population to be numbered, which makes this method less feasible if the number of individuals is huge.

24.2.1.2 Calculation of sample means and proportions

With simple random sampling, the calculation of sample mean and proportion is simple. See the first 10 chapters of this book.

24.2.1.3 Calculation of standard errors

As we mentioned before, the standard error of sample mean and proportion for a finite population would be calculated differently from that for an infinite population.

If n denotes the sample size, N the total number of individuals in a finite population, S standard deviation, p sample proportion, and $q = 1 - p$, then for an infinite population the standard error of sample mean is

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} \quad (24.1)$$

and the standard error of sample proportion is

$$S_p = \sqrt{\frac{pq}{n-1}} \approx \sqrt{\frac{pq}{n}}. \quad (24.2)$$

For a finite population the standard error of sample mean is

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} \times \sqrt{1 - \frac{n}{N}} \quad (24.3)$$

and the standard error of sample proportion is

$$S_p = \sqrt{\frac{pq}{n-1}} \times \sqrt{1 - \frac{n}{N}}. \quad (24.4)$$

In the above example, we are taking a sample of 300 students from a population of 3000 students, which is a finite population. Therefore, we should use formula (24.4) to calculate the standard error of the sample proportion. If the infection proportion is 0.1, the standard error of the sample proportion would be

$$S_p = \sqrt{\frac{0.1(1-0.1)}{300-1}} \times \sqrt{1 - \frac{300}{3000}} = 0.01046 \text{ or } 1.05\%.$$

All the populations we discussed in chaps. 1–10 of this book are infinite populations, and the readers should already be familiar with formulas (24.1) and (24.2). Notice that they can also be considered as the limits of formulas (24.3) and (24.4) when $N \rightarrow \infty$.

24.2.1.4 *Population parameters and confidence intervals*

Population parameters are statistical indicators for a specified population. We use μ to represent the mean of a population, and π to represent the probability of a population. We can use sample mean or proportion and their standard error to estimate the confidence interval of population mean or proportion. For the confidence interval of the mean of normal distribution, refer to Sec. 3.3 of chap. 3.

For the population probability π of a binomial distribution, if it is neither close to 0% nor close to 100.0%, and the sample size is sufficiently large, the confidence intervals can be estimated by a normal distribution approximately introduced in Chap. 3. When the population probability π is close to either 0 or 100.0%, or the sample size is less than 50, the confidence interval can be calculated directly by means of the theory of binomial distribution, which we have discussed in Sec. 2.3 of Chap. 2. Since this calculation can be quite complicated, one can refer to Table 3 of Appendix II.

24.2.2 *Systematic sampling (mechanical sampling)*

24.2.2.1 *Sampling method*

With this method, observational units are selected mechanically with a fixed interval according to certain order.

For example, to find the patients' degree of satisfaction with an outpatient department, we plan to take one tenth of the patients seeing the doctors as our sample. The systematic sampling is used to select one for every 10 outpatients for the survey. Before carrying out the research we randomly select a digit from 0 to 9 (e.g. 6) as a starting number. Then the 6th, 16th, 26th, 36th, etc. Are selected in sequence as the subjects of the sample.

It is easy to conduct Systematic sample. In general, there is less sampling error in systematic sampling than in simple random sampling.

24.2.2.2 *Calculation of standard error*

The sampling error of systematic sampling varies with the characteristics of the population and the length of the sampling interval, and the formula is rather complicated. Since the error is usually smaller than that in a simple random sampling, the formula for simple random sampling is often used

as a substitute. Of course, as a result, the estimated standard error is larger than the naïve one.

24.2.2.3 *Estimation of confidence intervals of population parameters*

The estimation methods are the same as those in simple random sampling.

24.2.3 *Stratified sampling*

24.2.3.1 *Sampling method*

When the stratified sampling method is used, the population should firstly be divided into several parts or blocks, called strata, according to certain characteristics that may have influence on the population parameters; then a simple random sampling is carried out within each stratum. In the same example of roundworm infection, a number of students could be randomly taken out from each of the grades. There are two ways to do this. One way, called proportionate stratified random sampling, is that the same proportion of subjects are taken from every stratum, e.g. 15% of students are taken from each grade; the another, called optimal assignment stratified random sampling, is that the sample size in each stratum is determined by the formula (24.34) or (24.36) in the following section, which will result in a smallest sample standard error.

When a population is divided into strata, they should be exhaustive and mutually exclusive; the differences among different strata should be as large as possible, and the differences among the individuals within each stratum should be as small as possible. If the proportionate stratified random sampling is chosen, the total number of individuals in each stratum should be known; if the optimal stratified random sampling is chosen, the standard deviations within each stratum should also be known.

24.2.3.2 *Calculation of sample mean and sample proportion*

Suppose there are k strata, the sample mean and the total number of individuals of the i th stratum are \bar{X}_i and N_i respectively, $i = 1, 2, \dots, k$, $N_1 + N_2 + \dots + N_k = N$. Then the formula to calculate sample mean is

$$\bar{X} = \sum_{i=1}^k \frac{N_i}{N} \bar{X}_i \quad (24.5)$$

and the formula to calculate sample proportion p is

$$p = \sum_{i=1}^k \frac{N_i}{N} p_i. \quad (24.6)$$

24.2.3.3 Calculation of standard errors

(1) For stratified random sampling For an infinite population, the standard error of sample mean is

$$S_{\bar{x}} = \sqrt{\sum_{i=1}^k \left(\frac{N_i}{N} \right)^2 \frac{S_i^2}{n_i}}, \quad (24.7)$$

where n_i and S_i^2 , $i = 1, 2, \dots, k$ are the sample size and sample variance in the i th stratum; the standard error of sample proportion is

$$S_p = \sqrt{\sum_{i=1}^k \left(\frac{N_i}{N} \right)^2 \left(\frac{p_i q_i}{n_i} \right)}, \quad (24.8)$$

where p_i and $q_i = 1 - p_i$, $i = 1, 2, \dots, k$ are the sample proportions in the i th stratum.

For a finite population, the standard error of sample mean is

$$S_{\bar{x}} = \sqrt{\sum_{i=1}^k \left(\frac{N_i}{N} \right)^2 \frac{S_i^2}{n_i} \left(1 - \frac{n_i}{N_i} \right)}, \quad (24.9)$$

the standard error of sample proportion is

$$S_p = \sqrt{\sum_{i=1}^k \left(\frac{N_i}{N} \right)^2 \left(\frac{p_i q_i}{n_i} \right) \left(1 - \frac{n_i}{N_i} \right)}. \quad (24.10)$$

(2) For proportionate sampling of an infinite population, the standard error of sample mean is

$$S_{\bar{x}} = \sqrt{\sum_{i=1}^k S_i^2 \frac{N_i}{Nn}}, \quad (24.11)$$

where n is the total sample size, $n = n_1 + n_2 + \dots + n_k$; the standard error of sample proportion from is

$$S_p = \sqrt{\sum_{i=1}^k p_i q_i \frac{N_i}{Nn}}. \quad (24.12)$$

For a finite population, the standard error of sample mean is

$$S_{\bar{x}} = \sqrt{\sum_{i=1}^k S_i^2 \frac{N_i}{Nn} \left(1 - \frac{n}{N}\right)}, \quad (24.13)$$

the standard error of sample proportion is

$$S_p = \sqrt{\sum_{i=1}^k p_i q_i \frac{N_i}{Nn} \left(1 - \frac{n}{N}\right)}. \quad (24.14)$$

24.2.3.4 Estimation of population parameters and confidence intervals

The same methods are used as those for simple random sampling.

24.2.4 Cluster sampling

24.2.4.1 Sampling methods

k clusters or groups are randomly selected from a population with K clusters or groups, and all the individuals or units in the selected clusters are taken as our sample. For instance, in a study on the prevalence of myopia among middle school students in an area, $k = 2$ schools are randomly selected from $K = 4$ schools, and all the students in the selected schools are surveyed. This is a cluster sampling. If two classes are further selected from each of the two schools, and all students in those two classes are surveyed, then this is a two-stage cluster sampling. Since it is relatively easy to perform, cluster sampling is widely adopted in large scale surveys. However, sampling errors can be quite large due to the large differences that might exist among the clusters.

24.2.4.2 Calculation of sample mean and proportion

The formula of sample mean is

$$\bar{x} = \frac{\sum X}{\sum m_i}, \quad (24.15)$$

where $\sum X$ is the summation of all observations in the sampled clusters. m_i is the number of individuals in cluster i and $\sum m_i$ is the summation of all individuals in the sampled clusters.

The formula of sample proportion is

$$p = \frac{\sum a_i}{\sum m_i}, \quad (24.16)$$

where $\sum a_i$ is the summation of all positive individuals in sampled clusters.

24.2.4.3 Calculation of standard errors

Suppose \bar{x}_i is the sample mean of cluster i , p_i is positive proportion of cluster i , and $\bar{m} = \sum m_i / k$.

(1) If the numbers of individuals in each sampled cluster are the same For an infinite population, the standard error of sample mean is

$$S_{\bar{x}} = \sqrt{\frac{\sum (\bar{x}_i - \bar{x})^2}{k(k-1)}} \quad (24.17)$$

and the standard error of sample proportion is

$$S_p = \sqrt{\frac{\sum (p_i - p)^2}{k(k-1)}}. \quad (24.18)$$

For a finite population, the standard error of sample mean is

$$S_{\bar{x}} = \sqrt{\frac{\sum (\bar{x}_i - \bar{x})^2}{k(k-1)} \left(1 - \frac{k}{K}\right)} \quad (24.19)$$

and the standard error of sample proportion is

$$S_p = \sqrt{\frac{\sum (p_i - p)^2}{k(k-1)} \left(1 - \frac{k}{K}\right)}. \quad (24.20)$$

(2) If the numbers of individuals in each sampled cluster are not the same
For an infinite population, the standard error of sample mean is

$$S_{\bar{x}} = \sqrt{\frac{\sum m_i^2 (\bar{x}_i - \bar{x})^2}{\bar{m}^2 k (k-1)}} \quad (24.21)$$

and the standard error of sample proportion is

$$S_p = \sqrt{\frac{\sum m_i^2 (\bar{p}_i - p)^2}{\bar{m}^2 k (k-1)}}. \quad (24.22)$$

For a finite population, the standard error of sample mean is

$$S_{\bar{x}} = \sqrt{\frac{\sum m_i^2 (\bar{x}_i - \bar{x})^2}{\bar{m}^2 k (k-1)} \left(1 - \frac{k}{K}\right)} \quad (24.23)$$

and the standard error of sample proportion is

$$S_p = \sqrt{\frac{\sum m_i^2 (\bar{p}_i - p)^2}{\bar{m}^2 k (k-1)} \left(1 - \frac{k}{K}\right)}. \quad (24.24)$$

Multi-stage cluster random sampling is widely used in cross-sectional studies, and sometimes it is followed by a stratified sampling. For example, to study the prevalence of hypertension among the residents in a large area, several countries in the area are selected; then several towns or villages within each country are selected; finally several communities or neighborhoods are selected; and all the residents living in those selected communities or neighborhoods become our subjects.

24.3 Estimation of Sample Size

In a cross-sectional study, it is often necessary to make sure that the estimation of a parameter is within a tolerance error and to know the effective minimum sample size. As stated in the previous text, three quantities must be predetermined in order to estimate the sample size.

First, the maximal tolerable error δ , that is, the allowable maximal difference between sample statistics and population parameter should be

predetermined. It will usually be half of the width of the desired confidence interval.

Second, the standard deviation σ or probability π of the population in question should be predetermined, which can be obtained from existing knowledge or by doing pilot studies.

Third, the confidence level $1 - \alpha$ should be predetermined, which is usually taken as $\alpha = 0.05$. In principle, the smaller the α is, the larger the sample size will be.

A desired sample size can be obtained either by looking up specific tables or by using certain formulas. Although tables are convenient to use, they have certain limitations. Formulas, which are very widely used, may vary according to what sampling method you are using. Generally speaking, cluster sampling results in the largest standard error, simple random sampling follows, then comes systematic sampling, and stratified sampling produces the smallest standard error. Thus, the sample size needed for simple random sampling will be less than that needed for cluster sampling and more than that needed for systematic and stratified sampling. There is no single formula for the sample size of systematic sampling as different sampling intervals result in different standard errors.

In this section, we will discuss how to estimate the sample size for cluster sampling, simple random sampling, and stratified sampling.

24.3.1 Sample size in cluster sampling

24.3.1.1 Sample size needed for estimating population proportion

In formula (24.25) below, k_0 is the number of clusters selected from an infinite population, k_y is the observed number of clusters in pilot study, m_i and p_i are the observed number of individuals and the proportion of positive events respectively in cluster i , which is obtained from a pilot study; δ is the tolerance error; Z_α is the two-tail critical value of standard normal distribution and it is commonly taken as $\alpha = 0.05$.

For an infinite population, the sample size needed for estimating population proportion is

$$k_0 = Z_\alpha^2 \sum \frac{m_i^2 (p_i - p)^2}{(k_y - 1) \bar{m}^2 \delta^2}. \quad (24.25)$$

For a finite population, the sample size needed for estimating population proportion is

$$k_1 = k_0 \left(1 - \frac{k_0}{K} \right). \quad (24.26)$$

In formula (24.26), k_1 is the number of clusters which should be selected from the population, k_0 is calculated by formula (24.25), and K is the total number in population.

24.3.1.2 Sample size needed for estimating population mean

For an infinite population, the sample size is

$$k_0 = Z_\alpha^2 \sum \frac{m_i^2 (\bar{x}_i - \bar{x})^2}{(k_y - 1) \bar{m}^2 \delta^2}. \quad (24.27)$$

For a finite population, the sample size is

$$k_1 = k_0 \left(1 - \frac{k_0}{K} \right). \quad (24.28)$$

In (24.27) and (24.28), meanings of $k_0, k_1, K, k_y, \sigma, \alpha, Z_\alpha, m_i, \bar{m}$ are the same as in formulas (24.25) and (24.26). \bar{x}_i is the sample mean in cluster i which is obtained from a pilot study, \bar{x} is the mean of k_y clusters, Σ is to take a summation over all clusters.

Example 24.1 Cluster sampling is going to be used in a study to investigate the prevalence of hypertension among people aged 40 and above in a city that has 55 communities. A pilot study has been done in two randomly selected communities, that 1060 cases among 4180 individuals are observed in the first community and 720 cases among 4970 individuals are observed in the second community; the hypertension proportion is 0.2536 and 0.1449 respectively. How many clusters should be taken as the sample to meet the need of the study ($\alpha = 0.05, \delta = 0.1$)?

Solution For this example, two-tail critical value of standard normal distribution is $Z_{0.05} = 1.96$, the tolerance error is $\delta = 0.1$, $\bar{m} = (4180 + 4970)/2 = 4575$, $p = (1060 + 720)/(4180 + 4970) = 0.1945$, $k_y = 2$,

and $K = 55$. By using Eq. (24.25) we have

$$k_0 = (1.96)^2 \frac{(4180)^2(0.2536 - 0.1945)^2 + (4970)^2(0.1449 - 0.1945)^2}{(2 - 1)(4575)^2(0.1)^2} \\ = 2.19 \approx 3.$$

As a result, three communities would be needed if the city had infinite clusters. But in this example the city serves as a finite population with $K = 55$ clusters. Therefore, the formula (24.26) must be used to adjust k_0 . The adjusted number of sampled clusters is

$$k_1 = 3 \left(1 - \frac{3}{55} \right) = 2.84 \approx 3.$$

That is, three communities are needed for the study.

24.3.2 Sample size in simple random sampling

24.3.2.1 Sample size needed for infinite population

As we have already known, if the probability is within a range of 0.2–0.8, the sample size needed to estimate the population probability is

$$n_0 = Z_\alpha^2 \frac{p_0(1 - p_0)}{(p - p_0)^2}, \quad (24.29)$$

where p is the sample proportion, p_0 is the population probability, $p - p_0$ is the tolerance error, and Z_α is the two-tail critical value of the standard normal distribution.

If the probability is smaller than 0.2 or greater than 0.8, a square root transformation is needed for data of proportion (in decimal fraction), where angles are expressed as radians. The sample size is calculated by

$$n_0 = \frac{Z_\alpha^2}{4 (\sin^{-1} \sqrt{p} - \sin^{-1} \sqrt{p_0})^2}. \quad (24.30)$$

When the population mean is estimated, the sample size is calculated by the following formula:

$$n_0 = Z_\alpha^2 \frac{\sigma^2}{\delta^2}. \quad (24.31)$$

24.3.2.2 Sample size needed for finite population

Sample size needed for a finite population is given by

$$n_1 = n_0 \left(1 - \frac{n_0}{N} \right), \quad (24.32)$$

where n_0 is calculated from formula (24.29), (24.30) or (24.31), N is the number of units or individuals of the finite population.

Example 24.2 A survey is planned to find the prevalence rate of myopia among sixth grade students of elementary schools. The observed rate was 8.0% from a pilot study done in 1992. The expected rate is about 10.0%. How many students are required given $\alpha = 0.05$?

Solution $p = 0.08$, $p_0 = 0.1$, $\alpha = 0.05$, $Z_{0.05} = 1.96$, $\sin^{-1} \sqrt{0.1} = 0.321751$, $\sin^{-1} \sqrt{0.08} = 0.286757$, using formula (24.30) we have

$$n_0 = \frac{(1.96)^2}{4 \left(\sin^{-1} \sqrt{0.1} - \sin^{-1} \sqrt{0.08} \right)^2} \approx 784.$$

That is, a sample of 784 students is needed.

Example 24.3 We want to investigate the mean level of hemoglobin of healthy adults in an area, with an error of not more than 0.2(g/l). The standard deviation is about 1.5(g/l) according to literature. How many subjects are needed with $\alpha = 0.05$?

Solution Given two-tail critical value $Z_{0.05} = 1.96$, $\delta = 0.2$, $\sigma = 1.5$, using Eq. (24.31) we have the sample size required given by

$$n_0 = (1.96)^2 \frac{(1.5)^2}{(0.2)^2} = 216.1 \approx 217.$$

That is, a random sample with 217 healthy adults is needed for the study.

24.3.3 Sample size in stratified sampling

24.3.3.1 Sample size needed to estimate population probability

This is given by

$$n = \frac{\left(\sum W_i \sqrt{p_i q_i} \right)^2}{V + \sum \frac{W_i p_i q_i}{N}}, \quad (24.33)$$

where $W_i = N_i/N$; N_i , p_i and q_i are the number of individuals, the proportion of positive events, and the proportion of negative events with $q_i = 1 - p_i$ for stratum i respectively; N is the total number of population; V is the estimated variance of the proportion, it is common to take $V = (\delta/Z_\alpha)^2$; δ is the tolerance error; Z_α is the two-tail critical value of standard normal distribution.

The estimated sample size n is then allocated into strata. The sample size n_i of stratum i is estimated by

$$n_i = \frac{n N_i \sqrt{p_i q_i}}{\sum N_i \sqrt{p_i q_i}} \quad (24.34)$$

Here the meanings of all notations are the same as above. Example is omitted.

24.3.3.2 Sample size needed to estimate population mean

This is given by formula

$$n = \frac{\sum W_i^2 S_i^2 / w_i}{V + \sum W_i S_i^2 / N} \quad (24.35)$$

where $W_i = N_i/N$, $w_i = N_i S_i / \sum N_i S_i$, N_i is the number of units in stratum i , S_i^2 is the sample variance of stratum i , N is the total number of units in the population, V is the required variance taken as $V = (\delta/Z_\alpha)^2$ in common, δ is tolerance error, Z_α is two-tail critical value of standard normal distribution.

The estimated sample size n is then allocated into the selected strata. The sample size n_i needed in stratum i is given by

$$n_i = \frac{n N_i S_i}{\sum N_i S_i} = n \times w_i \quad (24.36)$$

Example 24.4 A research is designed to evaluate the current mean weights of boys aged 2 to 4 years in a city by stratified sampling. Table 24.1 shows the mean weights and standard deviations of boys aged 2 to 4 years in this city measured in 1990. How many boys are needed for the study given $\alpha = 0.05$ and $\delta = 0.2$ kg? How many boys are needed for each age group?

Table 24.1 Sample size for the study on body weight (kg) of boys (stratified sampling).

Age <i>i</i>	Number <i>N_i</i>		Mean <i>X_i</i>	SD <i>S_i</i>	<i>N_i S_i</i> (2) × (5)	<i>w_i</i> (6)/∑ (6)	<i>W_i² S_i² / w_i</i>	<i>W_i S_i²</i>	<i>n_i</i> <i>n</i> × (7)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
2–	3773	0.2791	11.51	1.48	5584.04	0.2542	0.6712	0.6113	64
3–	5324	0.3938	13.93	1.89	10062.36	0.4580	1.2095	1.4067	115
4–	4422	0.3271	15.60	1.43	6323.46	0.2878	0.7602	0.6689	72
Total	13519	1.0000			21969.86	1.0000	2.6409	2.6869	251

Solution In this example $Z_{0.05} = 1.96$, $\delta = 0.2$, $V = (0.2/1.96)^2 = 0.0104$. Using formula (24.35), the total sample size n is given by

$$n = \frac{\sum W_i^2 s_i^2 / w_i}{V + \sum W_i s_i^2 / N} = \frac{2.6409}{0.0104 + \frac{2.6860}{13519}} = 249.26 = 250.$$

That is, a total of 250 boys are needed.

The number of boys needed in each age group n_i can be found in column 10 of Table 24.1, which is computed with formula (24.36). The sum of them is 251, slightly different from 250 due to rounding up.

24.4 The Current Life Table

There are two major types of life tables: the cohort life table and the current life table.

A cohort life table records the actual mortality experience of a particular group of individuals from the birth of the first member to the death of the last member of the group. It is useful in the analysis of prospective studies and clinical follow-up trials.

A current life table is formed on the basis of age-specific death rates prevailing in a particular population in certain period of time (say, certain year) obtained from a cross-sectional study. It is widely used in calculating life expectancies. As life expectancies based on the current life table are not affected by the age structure of a population, and the life expectancies of different populations can easily be compared. Let us suppose that a hypothetical generation of individuals were born at the same time and died later

on following the current age-specific death rate until the whole generation disappears. We can use the method of current life table to compute several life indices such as age-specific probability of death, the number of people dying, the number of people surviving longer than a certain age, and life expectancy, etc. These indices reflect the life process of a hypothetical generation of individuals with a series of given age-specific death rates.

A current life table may be a complete one or abridged one. In a complete life table the functions are computed for each year of life; whereas an abridged life table deals with age intervals greater than one year except for the first year of life. A typical set of intervals is 0 to less than 1 year of infant period, 1 to less than 5 years and every 5 years there after, i.e., 5 to less than 10 years, 10 to less than 15 years, and so on.

In the current life table x represents the starting age and n represents age span.

24.4.1 Construction and formulas of abridged current life table

Table 24.3 shows the format of abridged current life table. The columns are explained below:

24.4.1.1 Age-specific death rate (column 4)

It is the average death rate of the individuals in an age interval of one year. We can calculate it with the following equation

$$m_x = \frac{D_x}{P_x}, \quad (24.37)$$

where D_x is the number of deaths and P_x is the average population for the age interval $(x, x + n)$.

In theory, P_x should be the mid-year population, which is difficult to obtain. Therefore, an averaged population, (population at beginning + population at end)/2, is usually used instead. Because the death rate of the age group 0–1 varies greatly, the corresponding cell of the column is often left blank.

24.4.1.2 Age-specific probability of death q_x (column 5)

It is the conditional probability of a person alive at age x dying in the subsequent n years. Survival probability of a person, who is alive at age x ,

still survives during the next n years, is

$$p_x = 1 - q_x. \quad (24.38)$$

There is a functional relationship between the age-specific probability of death and the age-specific death rate. When the age interval is small, the functional relationship can be expressed as:

$$q_x = \frac{2nm_x}{2 + nm_x}. \quad (24.39)$$

As an exception, for the age group 0–1, we can use the death rate of infants as an estimate of the death probability.

24.4.1.3 *Number of survivors, l_x (column 6) and number of deaths, d_x (column 7)*

Let us assume that there were l_0 individuals born at the same time. l_0 is called the life table population, which can be any number set by the question of interest. In column 6, it is given that $l_0 = 100,000$. l_x denotes the number of individuals still alive at age x . d_x is the number of individuals who died during the age interval $(x, x + n)$. The relationship among l_x , d_x and q_x is

$$d_x = l_x q_x, \quad (24.40)$$

$$l_{x+n} = l_x - d_x. \quad (24.41)$$

24.4.1.4 *Number of person-years lived in the interval $(x, x + n)$ L_x (column 8)*

This number is contributed by the individuals alive exactly at x and is given by

$$L_x = n \left(\frac{l_x + l_{x+n}}{2} \right). \quad (24.42)$$

Number of alive person-years for infant group, L_0 , can be computed by

$$L_0 = l_1 + a_0 d_0, \quad (24.43)$$

where a_0 is the average fraction of the year lived by infants who die during the first year of life. Value of a_0 may be obtained from Table 24.2 provided

Table 24.2 Death rates of infants and corresponding a_0 values.

Infant's death rate (%)	< 20	20–	40–	60–
a_0	0.09	0.15	0.23	0.30

by World Health Organization (WHO) for infant death rate and homologous a_0 values. These are empirical constants.

Number of person-years lived by individuals aged 80 years and over is given by equation

$$L_{80+} = \frac{l_{80}}{m_{80+}}. \quad (24.44)$$

24.4.1.5 Total number of person-years lived beyond age x , T_x (column 9)

This is the sum from x to the oldest age group,

$$T_x = \sum L_x = L_x + T_{x+n}. \quad (24.45)$$

24.4.1.6 Life expectancy at age x , e_x (column 10)

It is the average years to be lived by a person at age x . Since the total number of years of life remaining to the l_x individuals is T_x , the estimate of life expectancy at age x is

$$e_x = \frac{T_x}{l_x}. \quad (24.46)$$

Example 24.5 In an area, the numbers of residents and deaths by age group for males are listed in columns 1 to 3 of Table 24.3. To work out an abridged current life table on the basis of the data.

Solution The results computed with the above equations are listed in columns 4 to 10 of the table.

m_x (column 4) denotes the average death rate of age interval $(x, x + n)$, which is computed by Eq. (24.37) with 6 decimal places. For example m_1 is calculated as

$$m_1 = \frac{D_1}{P_1} = \frac{841}{207327} = 0.004056.$$

Table 24.3 Abridged life table for males living in a city, in 1990.

Age group $x-$ (1)	Average population P_x (2)	Number of deaths D_x (3)	Death rate m_x (4)	Probability of death in ($x, x+n$) q_x (5)	Number of alive at x l_x (6)	Number of deaths in ($x, x+n$) d_x (7)	Person- years lived in ($x, x+n$) L_x (8)	Person- years lived beyond x T_x (9)	Life expectancy e_x (10)
0-	52087	2531	—	0.048592	100000	4859	96259	6772352	67.7235
1-	207327	841	0.004056	0.016095	95141	1531	377501	6676093	70.171(0)6
5-	428534	523	0.001220	0.006084	93610	569	466624	6298593	67.2858
10-	502742	391	0.000778	0.003881	93040	361	464297	5831969	62.6824
15-	437832	423	0.000966	0.004819	92679	447	462278	5367671	57.9168
20-	296355	392	0.001323	0.006592	92232	608	459642	4905393	53.1852
25-	413410	563	0.001362	0.006786	91624	622	456567	4445751	48.5215
30-	311755	502	0.001610	0.008019	91003	730	453188	3989184	43.8360
35-	249108	557	0.002236	0.011118	90273	1004	448855	3535996	39.1701
40-	230522	725	0.003145	0.015603	89269	1393	442864	3087141	34.5824
45-	207893	982	0.004724	0.023342	87876	2051	434254	2644277	30.0909
50-	185145	1457	0.007870	0.038588	85825	3312	420846	2210023	25.7500
55-	144344	1747	0.012103	0.058738	82513	4847	400450	1789177	21.6830
60-	115751	2375	0.020518	0.097585	77667	7579	369385	1388727	17.8800
65-	84700	2707	0.031960	0.147976	70088	10371	324509	1019342	14.5430
70-	55797	2748	0.049250	0.219254	59716	13093	265849	694832	11.6356
75-	33289	2418	0.072637	0.307369	46623	14330	197290	428984	9.2011
80-	20893	2912	0.139377	1.000000	32293	32293	231694	231694	7.1748

The value 0.048592 for q_0 (column 5) is the actual infant's death probability by other investigation, rather than by calculation with the above equation. All other q_x are computed by Eq. (24.39), such as

$$q_5 = \frac{2(5)(0.001220)}{2 + 5(0.001220)} = 0.006084$$

l_x (column 6) and d_x (column 7), except d_0 , are computed by Eq. (24.40) and (24.41) respectively. Suppose that a hypothetical 100,000 individuals were born at the same time, d_0 is computed by Eq. (24.40),

$$d_0 = l_0 q_0 = 100000 \times 0.048592 = 4859.$$

According to Eq. (24.41), we have

$$l_1 = l_0 - d_0 = 100000 - 4859 = 95141,$$

$$d_1 = l_1 q_1 = 95141 \times 0.016095 = 1531,$$

$$l_5 = l_1 - d_1 = 95141 - 1531 = 93610,$$

$$d_5 = l_5 q_5 = 93610 \times 0.006084 = 570.$$

Others are computed in the similar way.

L_x (column 8) is computed by Eq. (24.42) – (24.44). Infant death rate is 48.592%, and a_0 is 0.23 from Table 24.2. By using Eq. (24.43), we have

$$L_0 = l_1 + a_0 d_0 = 95141 + 0.23(4859) = 96259.$$

For all other age groups, L_x is computed by Eq. (24.42) as

$$L_5 = 5(l_5 + l_{10})/2 = 5(93610 + 93040)/2 = 466624.$$

Others are computed similarly.

L_{80+} , number of person-years lived at age 80 and over, is computed by Eq. (24.44) as

$$L_{80+} = \frac{32293}{0.139377} = 231694.$$

T_x (column 9) is computed by Eq. (24.45) upwards as

$$T_{80} = 231694,$$

$$T_{75} = L_{75} + L_{80} = 197290 + 231694 = 428984,$$

$$T_{70} = L_{70} + T_{75} = 265848 + 428984 = 694832.$$

Others are computed similarly.

e_x (column 10) is the expectation of life at age x , computed with Eq. (24.46) as

$$e_0 = \frac{T_0}{l_0} = \frac{6772351}{100000} = 67.72,$$

$$e_1 = \frac{T_1}{l_1} = \frac{6676093}{95141} = 70.17.$$

Others are computed similarly.

24.4.2 Analysis of life table

In a life table, the probability of death, the number of survivors, the number of deaths and the life expectancy are the main indices available for analyzing and evaluating life status of a population.

24.4.2.1 Age-specific probability of death q_x

A U-shaped curvilinear diagram is displayed by drawing the age-specific death probability q_x of a life table on a semi-logarithmic scaled paper (Fig. 24.1). One can look at the height of the starting point of the curve, which reflects the magnitude of infant death rate; one can also look at the width of the bottom part of a curve and lowest point, of which the lowest

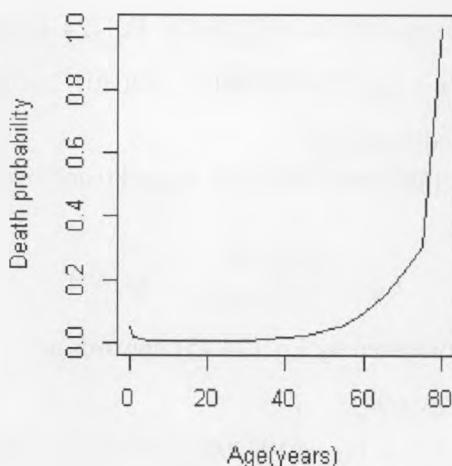


Fig. 24.1 Curve of age-specific death probability (male, 1981).

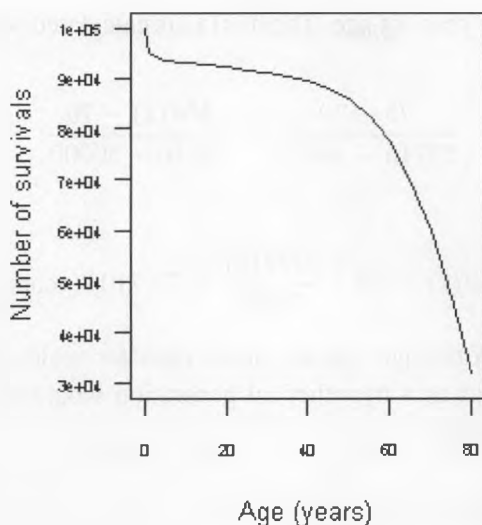


Fig. 24.2 Curve of life table survivals (male, 1981).

point usually occurs around the age group 10– before rising gradually; starting from age group 50–, the curve usually goes up sharply, which indicates the fast increasing death rates. Finally, we should pay attention to the tail of the curve: the more steep the slope is, the more rapidly the probability of death rises.

24.4.2.2 Number of survivals l_x

The number of survivals reflects the life process of a hypothetical generation of individuals experiencing current age-specific death rate. Usually a diagram is drawn and its height and curvature reflect the process of decreasing the number of survivals (Fig. 24.2). The lower the age-specific death rate is, the higher the curve is, and vice versa. The ratio between the numbers of two age groups, l_{x+n}/l_x , describes the health status in view of survivorship. The median survival age $Md(x)$ is used to describe the health status from another perspective, at which half of the population are expected to survive.

$Md(x)$ is estimated by interpolation. For example, in Table 24.3, we find the age group where survivors account for half of the hypothetical population (50,000). Since the age group 70– has 59,716 survivors, and the age group 75– has 46,623 survivors, the $Md(x)$ is located in the interval of

70 to less than 75 years of age. The $Md(x)$ is calculated with interpolation method as

$$\frac{75 - 70}{59716 - 46623} = \frac{Md(x) - 70}{59716 - 50000}.$$

We have

$$Md(x) = 70 + \frac{5(9716)}{13093} = 73.7103 \text{ (years)}.$$

That is, based on the age specific death rates for males in 1981, half of the male newborns of a hypothetical generation would have survived till 73.7103 years of age.

24.4.2.3 Number of deaths d_x

As oppose to the number of survivors, the number of deaths describes the death process of a hypothetical generation of individuals experiencing the current age-specific rates of death. The data can be used to draw histogram. The heights of the bars reflect the amount of deaths in each age group (Fig. 24.3).

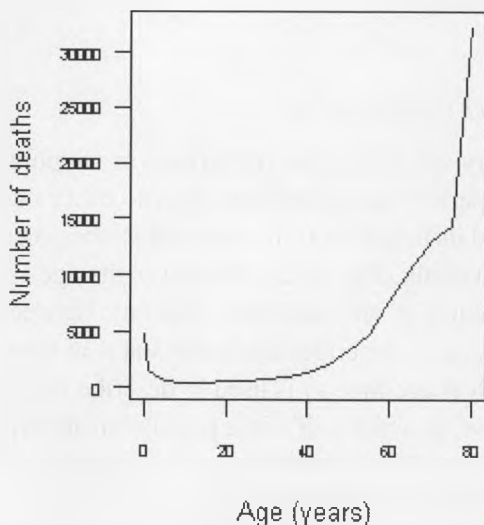


Fig. 24.3 Curve of death number by life-table (male, 1981).

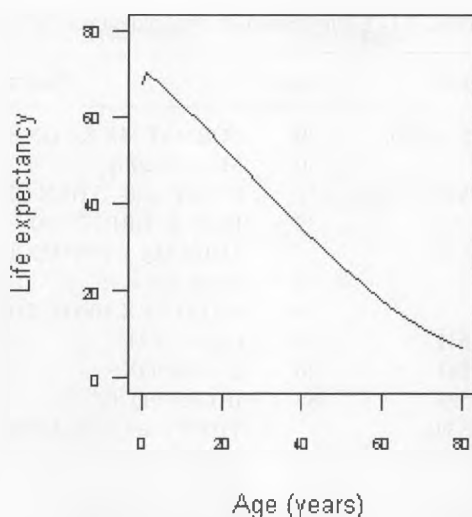


Fig. 24.4 Curve of life expectancy (male, 1981).

24.4.2.4 Life expectancy e_x

Life expectancy, or expectation of life, summarizes the mortality experience of people surviving beyond age x in the population. e_0 , or the life expectancy of newborns, reflects the average length of life that the individuals in the hypothetical generation born at the same time would live. e_0 is influenced by the mortality rates of all the age groups.

Life expectancy can be presented in a plot (Fig. 24.4). It is important to examine the starting point of the curve. If the age-specific death rate decreases, the starting point of the curve would be higher, and so does the curve as a whole. Since life expectancies are not affected by the age structure and can serve as a representative of the overall mortality rates of the whole population, those of different geographic areas, or different times can be compared side by side. e_x ($x \neq 0$) is the remaining years of life expected to be lived by a person at age x . It is a good reference in the field of insurance and social welfare.

24.5 Computerized Experiments

Experiment 24.1 Computation of abridged current life table The SAS code showed in Program 24.1 is to compute abridged current life table for data in Example 24.4.

Program 24.1 Computation of abridged current life table.

Line	Program	Line	Program
01	OPTIONS LS = 100;	29	FORMAT MX 8.6 QX 8.6;
02	DATA LIFE;	30	MX = DX/PX;
03	INPUT GRP\$ N PX DX;	31	IF GRP = '0-' THEN QX = DX/PX;
04	A0 = 0.23;	32	ELSE IF GRP NE '80-'
05	RETAIN I 0;		THEN QX = 2*N*MX/(2+N*MX);
06	I = I + 1;	33	ELSE QX = 1;
07	CARDS;	34	RETAIN LX 100000 D 0;
08	0- 1 52087 2531	35	LX = LX - D;
09	1- 4 207327 841	36	D = LX*QX;
10	5- 5 428534 523	37	IF GRP EQ '0-'
11	10- 5 502742 391		THEN L = LX - (1 - A0)*D;
12	15- 5 437832 423	38	IF GRP NE '80-'
13	20- 5 296355 392		THEN L = N*LX - 0.5*N*D;
14	25- 5 413410 563	39	ELSE L = LX/MX;
15	30- 5 311755 502	40	OUTPUT;
16	35- 5 249108 557	41	PROC SORT;
17	40- 5 230522 725	42	BY DESCENDING I;
18	45- 5 207893 982	43	DATA LIFE2;
19	50- 5 185145 1457	44	SET LIFE1;
20	55- 5 144344 1747	45	RETAIN TX 0;
21	60- 5 115751 2375	46	TX = TX + 1;
22	65- 5 84700 2707	47	EX = TX/LX;
23	70- 5 55797 2748	48	OUTPUT;
24	75- 5 33289 2418	49	PROC SORT;
25	80- 5 20893 2912	50	BY I;
26	;	51	PROC PRINT;
27	DATA LIFE1;	52	VAR GRP PX DX MX QX LX D L TX EX;
28	SET LIFE;	53	RUN;

In the above program, line 01 sets the width of the output by specifying $LS = 100$ in order to print a complete table; lines 02 to 26 are to create a new SAS data set LIFE; line 03 specifies the variables: GRP\$ represents the age groups, N represents the intervals of age groups, PX represents the averaged population by age group, and DX represents the actual numbers of deaths; line 04 gives a value of 0.23 for a_0 ; line 05 sets the initial value of I as 0 for each age group, and line 06 adds 1 to I when the age group goes up to the next one, in this case, I increases by 1 each time the age increases; lines 27–40 are to create a new dataset LIFE1 from LIFE; line 29 specifies the format of output for MX and QX as 8-digits numbers with

six decimal places; line 30 computes the death rate of each age group; lines 31–33 compute the death probabilities; line 34 gives the initial values $LX = 100000$ and $D = 0$; lines 35 and 36 are to compute the survivor numbers LX (i.e., l_x) and the death numbers D (i.e., d_x); lines 37–39 are to compute the numbers of person-years lived (i.e., L_x); line 40 is to output the results to data set LIFE1; lines 41–42 are to sort data by I in descending order; lines 43–44 are to create a data set LIFE2 from data set LIFE1; line 45 sets the initial value of total person-years lived beyond age x as $T_x = 0$; lines 46–47 are to compute the total person-years lived beyond age x , T_x and the life expectancy e_x ; line 48 outputs the final results to data set LIFE2; lines 49–50 are to resort I by ascending order; lines 51–53 print all the results of the life table. The results might be a slightly different from Table 24.3 due to rounding.

24.6 Practice and Experiments

1. Identify which sampling method is used in each of the following cases.
 - (1) A group of students are asked to draw lots, and two of them who get number 1 or 2 are selected.
 - (2) We randomly pick ten pages from a telephone book, and all the numbers listed on those pages are to be called during our telephone survey.
 - (3) 803 is a 3-digit number randomly generated by computer, and we pick those drivers whose driver license numbers end in 803.
 - (4) By rolling a dice, one soldier is picked out of six to carry out a task.
2. We want to study smoking habits of 2000 male workers in a factory. In a sample of 50 workers, 23 smokers have been found. Try to estimate the 95% confidence interval of the proportion of male workers who are smokers. If this is a pilot study, how many male workers would be needed if half length of the 95% confidence interval δ is required not to be more than 3%?
3. Myopia rates among middle school students in three school districts of a country, which has a total of 50 school districts, are measured to be 15%, 18% and 23%, and their total numbers of students are 180, 250 and 270 respectively. Try to estimate the 95% confidence interval of myopia prevalence in the country. If this is a pilot study, how many districts are

Table 24.4 Numbers of individuals and deaths by age group for a male population in 1993.

Age group	Averaged population	Number of deaths	Age group	Averaged population	Number of deaths
0	36813	517	45–	131043	534
1–	110489	143	50–	104152	738
5–	130692	117	55–	79966	1088
10–	191877	139	60–	60040	1581
15–	236564	202	65–	45111	1956
20–	225728	269	70–	23132	1814
25–	245295	269	75–	11004	1320
30–	195609	247	80–	3950	756
35–	132978	241	85–	1009	310
40–	119047	306			

required given half length of the 95% confidence interval δ less than or equal to 3%?

- Table 24.4 lists the numbers of individuals and deaths by age group for a male population in 1993. Construct an abridged current life table and report the life expectancy e_x .
- From Table 24.3, the life expectancy was 67.72 years at the age zero for male in 1981. It is suggested that the life expectancy would increase 5 years in the period of coming 5 years. Is it possible to reach the goal if all the death rates (m_x s) would decline 5.0% in 5 years? What about 10.0% or even 20.0% in 5 years? Make use of the SAS Program 24.1 to do this.

(1st edn. Songlin Yu; 2nd edn. Songlin Yu, Jiqian Fang)

Chapter 25

Design and Analysis of Prospective Studies

Prospective study is also called cohort study or follow-up study. In this type of study subjects exposed to different levels of possible etiological factors are followed over a period of time to observe who develop disease in question. The information is used to analyze the association between disease and exposure. Prospective approach involving looking forward from causes to effects is commonly adopted in clinical medicine, preventive medicine, and etiological studies. The main disadvantages of this kind of study are that large sample size and long period of observation are required. Therefore it needs more resources and expenditure, and it is difficult to follow up and prone to loss of subjects or censoring because of migration, secession from observation, and death from un-relevant disease, etc.

25.1 Study Design

For prospective study the exposure factor can be natural existence (e.g., smoking behavior or occupational exposure), or added by investigators (e.g., therapeutic drug in clinical trials or interfering measures in prevention medicine). As it is a kind of field study for which subjects are human beings, ethical issues should be taken note of in every step. The study should also follow statistical principles.

25.1.1 Study population

For source of subjects, the investigator should consider some special aspects like compliance, feasibility of communication, and completeness of medical

records of subjects, etc. There are three sources of subjects

(1) General population Prospective study is carried out based on community. The Framingham Heart Study is a good example of this type of study. It was initiated in 1948 by the United States Public Health Service in order to study the relationship of a variety of factors to the subsequent occurrence of heart disease. The area of Framingham, Massachusetts, was chosen for its population stability, incorporating prior studies, availability of a community hospital and proximity to a large medical center. The population in 30–60 age group was approximately 10,000 and a final sample of 5000 and more persons free from atherosclerosis heart disease were selected. After the first examination, each person was re-examined at two-year intervals for a 20-year long period. It was found that blood pressure, serum cholesterol, and cigarette smoking etc. were related to heart disease.

(2) Special population Doll and Hill's study on cigarette smoking and lung cancer is an example. In their study all physicians on the British Medical Register who were living in the United Kingdom were selected as subjects because they were much concerned with their own healthy status and maintained contact with several professional organizations. Information was available at the General Medical Council or British Medical Association.

(3) Hospitalized population Niswander and Gorden selected 5400 pregnant women as subjects, who had received antenatal examination and induced abortion in 12 cooperative hospitals. They were followed up until they delivered newborn during 1959–1966. The goal of the study was to evaluate the relation of prenatal mortality rate, infant mortality rate and morbidity rate of newborn to antenatal examination and induced abortion.

25.1.2 Control population

In order to determine the effect of exposure factor on disease occurrence, a control population is required as baseline for comparison. There are three types of control population.

(1) A group of people with no or lowest exposure may be selected from the same study population as control population. This type of control is called internal reference.

(2) Control individuals are selected from non-exposed population. They are comparable to exposed group with age, gender, and healthy status, etc.

(3) General population in the same area and the same time period as exposed subjects is selected as control. This type of control is called external reference.

In occupational epidemiology workers in duty are usually selected by passing through some special physical examination. Their health status may be better than, and their morbidity and mortality rates may be less than that of common population. Control group should be selected to be comparable to exposed group in order to avoid so-called "healthy worker effect" phenomenon.

25.2 Measures of Disease Occurrence

The frequency of a disease indicates the strength of a disease occurring in a population. There are two measures to describe frequency of a disease according to data resources: cumulative incidence probability and person-time incidence rate.

25.2.1 Cumulative incidence probability

25.2.1.1 Approximation of the Incidence probability

Let n denotes the size of total population followed up, and d denotes the number of new cases diagnosed during the study period. The incidence probability can be calculated approximately by

$$q = \frac{d}{n}. \quad (25.1)$$

This is a proportion of subjects, who develop the disease during the study period, to the population, who are disease-free at the beginning of the study. It reflects the possibility of disease occurrence for a person without disease previously. (25.1) can be used as an estimate of probability of disease occurrence provided n is large enough. For example, 1000 workers were exposed to the dust in work environment; in the period of 20 years afterwards, 200 new cases of silicosis were diagnosed. The incidence probability of silicosis

in 20 years is about

$$q = \frac{200}{1000} = 0.20 = 20\%.$$

For its ease in calculating and clear in meaning, the incidence probability is often available for data with short research period. If the loss in the number of persons follow up during the period is c , then (25.1) may be adjusted to

$$q = \frac{d}{n - \frac{c}{2}}. \quad (25.2)$$

25.2.1.2 *The method of cohort life table for cumulative incidence probability*

When research period is long, subjects enroll and withdraw so frequently at different time that the length of time followed-up for different individuals varies substantially and the probabilities of incidence may not keep constant for the whole period of follow-up, which means that Eqs. (25.1) and (25.2) cannot be used efficiently.

It is wise to divide the research period into several consecutive intervals $(k, k+1)$, $k = 0, 1, 2, \dots, m$, in each of which the incidence probability can be estimated by (25.1) or (25.2); the cumulative incidence probability can be obtained by integrating the incidence probabilities calculated in those intervals. This is just the basic idea of the method of cohort life table, which is also called actuarial method.

For the k th interval, let n_k be the number of subjects at the beginning of the interval, d_k be the number of patients occurring in the interval, c_k be the number of subjects lost to follow up in the interval and q_k be the incidence probability in the interval, which can be calculated by

$$q_k = \frac{d_k}{n_k - \frac{c_k}{2}}. \quad (25.3)$$

The probability of disease-free in the interval $(k, k+1)$ is $1 - q_k$. The cumulative probability of disease-free for a person over the m intervals is

$$p_{0,m} = \prod_{k=0}^{m-1} (1 - q_k).$$

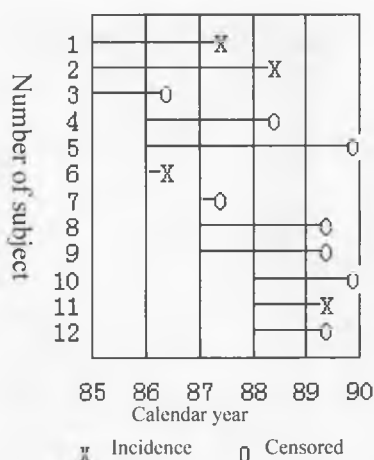


Fig. 25.1 Diagrammatic illustration of five year follow-up of 12 subjects by calendar year.

Then the cumulative incidence probability from time 0 to time m is

$$q_{0,m} = 1 - \prod_{k=0}^{m-1} (1 - q_k). \quad (25.4)$$

In fact, this is the estimate of probability for a person developing the disease over the whole study period.

Example 25.1 (Hypothetic) Consider a cohort study lasting for five years from the first day of 1985 to the end of 1989. 12 subjects are admitted into the cohort at the beginning of the corresponding year, and followed up until 1990. The results are plotted in Fig. 25.1.

For subjects 5 and 10, observations are cutoff at the end of 1989 because the study terminates. The two subjects are treated as censored. Figure 25.1 displays the original data by calendar year and Fig. 25.2 is resorted by observed years (end time–start time).

Solution Column 6 of Table 25.1 lists the incidence probabilities in the intervals calculated by Eq. (25.3). The rightmost column of the table lists cumulative incidence probabilities from $k = 0$ to $k = m$ by Eq. (25.4). The process from column 6 to column 7 is demonstrated in Table 25.1a. It is showed that cumulative incidence probability is a non-decreasing function. It goes up as observational time extends.

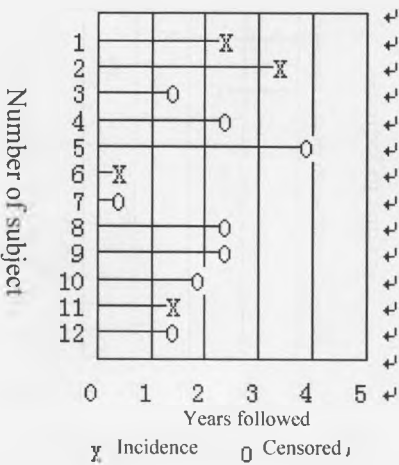


Fig. 25.2 Diagrammatic illustration of five year follow-up of 12 subjects by observation years.

Table 25.1 Calculations of interval and cumulative incidences of a cohort with 12 subjects.

Time interval $k \sim k + 1$ (1)	Number of cases in interval d_k (2)	Number of censored in interval c_k (3)	Number of subjects free from disease at k n_k (4)	Adjusted $n'_k = \frac{n_k - c_k}{2}$ (5)	Incidence probability in interval q_k (6)	Cumulative incidence probability up to $k + 1$ $q_{0,k+1}$ (7)
0–	1	1	12	11.5	0.087	0.087
1–	1	2	10	9.0	0.111	0.188
2–	1	4	7	5.0	0.200	0.351
3–	1	0	2	2.0	0.500	0.675
4–	0	1	1	0.5	0.000	0.675

25.2.2 Person-time incidence rate

In etiological studies of chronic diseases, follow-up period may sustain for a long time such as several years, ten more years, or even decades. Furthermore, as the subjects enroll and withdraw quite often, the length of time being followed-up may be different from one another. These may affect the estimate of probabilities in the intervals calculated by Eq. (25.3). Person-time incidence rate is another choice of measures. If we use year as measurement unit of time, and thus call it person-year incidence rate

Table 25.1a The process from column 6 to column 7 in Table 25.1.

Time interval $k-k+1$ (1)	Incidence probability in interval q_k (6)	Probability of non-incidence in interval $p_k = 1 - q_k$ (6)'	Cumulative non-incidence probability up to $k+1$ $p_{0,k+1} = p_0 p_1 \cdots p_k$ (6)''	Cumulative incidence probability up to $k+1$ $q_{0,k+1} = 1 - p_{0,k+1}$ (7)
0-	0.087	0.913	0.913	0.087
1-	0.111	0.889	0.812	0.188
2-	0.200	0.800	0.649	0.351
3-	0.500	0.500	0.325	0.675
4-	0.000	1.000	0.325	0.675

instead. Person-year incidence rate f is calculated by

$$f = \frac{d}{T}, \quad (25.5)$$

where T is the amount of person-years observed, d is the number of disease occurrences.

If a subject has exposed for a year, he or she contributes 1 year to the denominator of (25.5). If a subject has exposed for ten years, he or she contributes ten years to the denominator. The rate is often multiplied by 10^3 or even 10^5 to keep the significant digits. It is called the rate per 10^3 person-years or the rate per 10^5 person-years. For example, the total of person-years contributed by 12 subjects in Example 25.1 is

$$T = 2.5 + 3.5 + \cdots + 1.5 + 1.5 = 25 \text{ (person-years)}$$

and the number of disease occurrences is $d = 4$. The incidence rate is $f = 4/25 = 0.16$ cases/person-year, or 160 cases per 1000 person-years. The result shows that during the 5 years, 160 new cases of the disease would occur for every 1000 person-years on average.

One can see that the person-year incidence rate is not simply a measurement of probability in nature. In epidemiology as well as medical statistics, it is subject to the concept of intensity. In fact,

Person-year incidence rate

$$\begin{aligned}
 &= \lim_{\Delta \rightarrow 0} \frac{P(\text{new case occurring in}(t, t + \Delta) \mid \text{disease-free at } t)}{\Delta} \\
 &\approx \frac{P(\text{new case occurring in}(t, t + \Delta) \mid \text{disease-free at } t)}{\Delta} \\
 &\approx \frac{\text{number of new cases occurring in}(t, t + \Delta)}{(\text{number of persons disease-free at } t)\Delta}.
 \end{aligned}$$

The last two approximations exist when Δ is small.

In a way similar to the calculation of incidence probability in a long period, the whole period can be divided into several consecutive short intervals; the person-year incidence rate can be calculated for each of these short intervals. Then the person-year incidence rate of each interval is translated into incidence probability under an assumption of exponential distribution (that is, assuming the k th person-year incidence rate keeps constant in the k th interval).

$$q_k = 1 - \exp(-f_k \Delta_k), \quad (25.6)$$

where Δ_k is the time span of the k th short interval.

This process is illustrated with Example 25.1 as follows:

Step 1: In k th interval, the person-year incidence rate f_k can be translated into the incidence probability q_k by the Eq. (25.6).

In Table 25.2 all time spans of the short intervals are 1 year. The incidence probabilities translated from the person-year incidence rates are listed in Column 6 of Table 25.2.

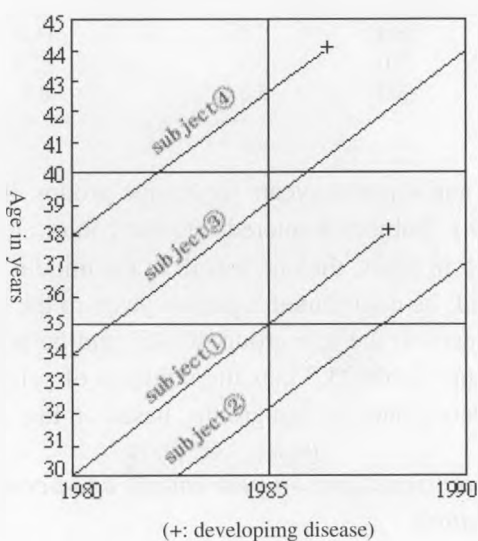
Step 2: Person-year cumulative incidence rate is obtained by using Eq. (25.4) similar to those in Table 25.1a. The results are listed in the rightmost column of Table 25.2.

25.2.3 Age-specific person-time rate

Many diseases are closely associated with age. In a long period of follow-up study, subjects are getting older as time increases as shown in Fig. 25.3. Subject 1 entered into the cohort at the beginning of 1980 when he was 30 years old and was ill in the middle of 1988 when he was 38.5 years old; he contributed 5 person-years to the age group 30–35 and 3.5 person-years to the age group 35–40; to the latter age group he became a new case. Subject 2

Table 25.2 Calculation of incidences for data of Example 25.1.

Time interval $k-k+1$	Number of subjects free from disease at k n_k	Number of cases occurred in interval d_k	Number of person-years followed T_k	Person-year incidence rate in interval f_k	Incidence probability in interval q_k	Cumulative incidence probability up to $k+1$ $q_{0,k+1}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0-	12	1	11.0	0.091	0.087	0.087
1-	10	1	8.5	0.118	0.111	0.188
2-	7	1	4.0	0.250	0.221	0.367
3-	2	1	1.5	0.667	0.487	0.675
4-	1	0	0	0.000	0.000	0.675

**Fig. 25.3** Ages of subjects are getting older as time increases.

entered into the cohort at the beginning of 1983 when he was 30 years old and ended the observation at the end of 1989 when he was 37 years old; he contributed 5 person-years to 30–35 age group and 2 person-years to the age group 35–40. But he was not a new case in either of the two age groups. Subject 3 entered into the cohort at the beginning of 1980 when he was 34 years old and ended the observation at the age of 44 years; he

Table 25.3 Person-year incidence rates by age groups from Fig. 25.3.

Age group	Number of person-years	Number of new cases	Rate per person-year
30–	11.0	0	0
35–	12.5	1	0.08
40–45	8.5	1	0.12
Total	32.0	2	0.06

Table 25.4 Person-years calculated by age-year group.

Age group	Calendar year		Summation of person-years
	1980–1985	1985–1990	
30–	8.0	3.0	11.0
35–	7.0	5.5	12.5
40–45	3.0	5.5	8.5

contributed 1, 5, and 4 person-years to the age groups 30–35, 35–40 and 40–45, respectively. Subject 4 entered into the cohort at the beginning of 1980 at the age of 38 years; then he was ill in the middle of 1986 when he was 44.5 years old; he contributed 2 person-years to the age group 35–40 and 4.5 person-years to the age group 40–45; and he was a new case to the latter age group. Table 25.3 lists the numbers of person-years and the person-year incidence rates by age groups, based on Fig. 25.3.

25.2.4 Two-dimensional person-year rate by age-period cross classification

The person-year rate can be calculated by age-period two-dimensional classification provided the number of subjects is large enough. For example, the numbers of person-years by age-year cross classification, from Fig. 25.3, are shown in Table 25.4. It is obvious that the summation by row is the number of person-years calculated by age group only, e.g., the numbers listed in the second column of Table 25.3. In the same way as Table 25.4, the number of developing disease in each cell can be calculated. Then the person-year incidence rate by age-year group can be calculated for every cell in a two-way table. In occupational epidemiology, as an example, work environment,

strength and duration of exposure differ in different time, etc. such that the person-year incidence rate changes frequently. The two-way classification method here can be applied to find the pattern of disease occurrence.

25.3 Analysis of Data from Prospective Studies

Prospective studies are used to investigate the etiology or treatment effect of a disease. In order to analyze the association between disease and exposure factor, it is necessary to compare disease frequencies of different exposed groups. The indices used for comparative analysis are relative risk (RR), attributable risk (AR), population attributable risk (PAR), and dose-response relationship, etc. When there are confounders in addition to the exposure factor, both stratified analysis and multivariate analysis are often used to control the confounding effects.

25.3.1 Relative risk

Relative risk is a ratio of disease frequency for exposed group (q_1 or f_1) divided by disease frequency for control or reference group, which is used as baseline (q_0 or f_0). $RR = 1.0$ means disease frequency for exposed group is the same as that for control group. $RR - 1$ means the increased (or decreased) fraction for exposed group than for control group. For example, $RR = 2.5$ means disease frequency for exposed group is 2.5 times as many as that for control group, e.g., the frequency for the former is $2.5 - 1 = 1.5$ times higher than that for control group; $RR = 0.60$ means the frequency for exposed group is 0.6 times as many as that for control group, and $0.6 - 1 = -0.4$ means frequency for exposed group is 0.4 times lower than that for control group.

The estimate of relative risk in terms of incidence probability is

$$\hat{RR} = \frac{q_1}{q_0}, \quad (25.7a)$$

where q_1 and q_0 are the incidence probabilities in the exposed group and non-exposed group respectively.

The estimate of relative risk in terms of person-year incidence rate is

$$\hat{RR} = \frac{f_1}{f_0}, \quad (25.7b)$$

Table 25.5 Data layout of prospective study.

Follow-up results	Factor		Total
	Exposed	Non-exposed	
Number of disease occurrences	a	b	m_1
Number of disease-free	c	d	m_0
Total	n_1	n_0	n

where f_1 and f_0 are the person-year incidence rates in exposed group and non-exposed group respectively.

In order to avoid the effect of sampling error on inference, it is necessary to work out a statistical test

$$H_0 : RR = 1.0, \quad H_1 : RR \neq 1.0.$$

We will introduce two methods with examples for illustration. The choice of methods is based on statistical characteristics of the data.

25.3.1.1 Mantel-Haenszel χ^2 test

It is applied to \hat{RR} calculated with cumulative incidence probabilities as (25.7a). The layout of data is shown Table 25.5. The Mantel-Haenszel statistic χ^2 is

$$\chi^2 = \frac{(n-1)(ad-bc)^2}{n_1 n_0 m_1 m_0}, \quad (25.8)$$

where the meaning of symbols are explained in Table 25.5. Obviously, χ^2 in Eq. (25.8) is similar to the statistic of χ^2 for the test in 2×2 table of Chap. 6, except the numerator, which is $(n-1)$ in Eq. (25.8) not n in Eq. (6.8a). When n is sufficiently large, both equations are equivalent.

Based on the knowledge introduced in Chap. 6, under H_0 , the statistic χ^2 is distributed as $\chi^2_{(1)}$ when n is large enough. According to the value of χ^2 a statistical decision can be made.

The 95% confidence interval of RR can be calculated by

$$\hat{RR} \left(1 \pm \frac{1.96}{\sqrt{\chi^2}} \right), \quad (25.9)$$

Table 25.6 Data from prospective study on association between CAT and CHD.

Follow-up results	Level of blood catecholamine		Total
	High	Low	
Occurrence of CHD	27	44	71
Non-occurrence of CHD	95	443	538
Total	122	487	609

Example 25.2 A prospective study was designed to investigate the association between the subsequent coronary heart disease (CHD) and the blood serum catecholamine (CAT) level. CAT was regarded as an exposure factor with two categories (high, low). 609 male adults had enrolled into the study and their blood CATs were examined. Then the subjects were divided into two groups according to their CAT levels. After seven years of follow-up, 71 new cases of CHD had been identified. The data are listed in Table 25.6.

The relative risk of high level blood CAT group comparing to low level group using (25.7a) is:

$$\hat{RR} = \frac{\frac{27}{122}}{\frac{44}{487}} = \frac{0.221}{0.08} = 2.45 \text{ (times).}$$

By using Eq. (25.8), when H_0 is true, the test statistic is

$$\chi^2 = \frac{(609 - 1)(27 \times 443 - 44 \times 95)^2}{122 \times 487 \times 71 \times 538} = 16.22.$$

Referring to the table of χ^2 distribution, the critical value with one degree of freedom is $\chi_{0.05}^2 = 3.84 < 16.22$. The null hypothesis is thus rejected and the incidence probabilities for the two groups are significantly different. The incidence probability for high level group is 2.45 times as many as that for low level group. The 95% confidence interval of RR using Eq. (25.9) is

$$2.45 \left(1 \pm \frac{1.96}{\sqrt{16.22}} \right) = (1.58, 3.79).$$

Table 25.7 Data layout of person-years for calculating rates.

Follow-up results	Factor		Total
	Exposed	Non-exposed	
Number of occurrences	a	b	m
Amount of person-years	T_1	T_0	T
Rate per person-year	f_1	f_0	f

25.3.1.2 Method based on binomial distribution and normal approximation

It is applied to \hat{RR} calculated with person-year incidence rates as (25.7b). The layout of the data is showed in Table 25.7. Under the two rates would be equal to each other. Let the amount of person-years for exposed and non-exposed groups be T_1 and T_0 , respectively. The total amount of person-years is $T = T_1 + T_0$. The proportion of person-years for exposed group in the total amount is $p = T_1/T$. The proportion of the disease occurrences for exposed group in the total disease occurrences $a/(a + b)$ should also be p under H_0 . We use p as the estimate of the probability of event occurrence. Then the number of disease occurrences follows a binomial distribution $B(m, p)$, where $m = (a + b)$. When exposure is a harmful factor, the alternative hypothesis is $H_1 : RR > 1$. Under the null hypothesis H_0 the probability $P(a \leq x \leq m)$ can be calculated by the law of binomial probability

$$P(a \leq x \leq m) = \sum_{x=a}^m \binom{m}{x} p^x (1 - p)^{m-x}, \quad (25.10)$$

where x is a random variable representing the number of disease occurrences in the exposed group.

If the probability is equal to or less than a , the null hypothesis H_0 is rejected and the relative risk is significantly greater than 1.0.

When m is large enough, the probability calculated by Eq. (25.10) can be approximated with a normal score as

$$Z = \frac{|a - mp| - 0.5}{\sqrt{mp(1 - p)}}. \quad (25.11)$$

Table 25.8 Data from a prospective study for possible association of obesity with death.

Results of follow-up	Body weight status		Total
	Obesity	Non-obesity	
Number of deaths	30	36	66
Amount of person-years	699	1399	2098
Rate per person-year	0.043	0.026	0.031

And then compare with a one-side critical value of the standard normal distribution.

When the exposure is a possible protective factor, the alternative hypothesis is $H_1 : RR < 1.0$. In this situation it is necessary to calculate the probability $P(0 \leq x \leq a)$. The only change in (25.10) is that the summation on the right-hand side should be from 0 to a for x , instead of from a to m . The normal approximation still keeps the same as (25.11).

Example 25.3 In a study of association between death and obesity, a number of women aged 60–75 years old were assigned into either obesity group or non-obesity group. After follow-up for 8 years, the numbers of deaths and person-years for the two groups were obtained as showed in Table 25.8.

By using Eq. (25.7b), the relative risk in terms of person-year incidence rates is $\hat{RR} = 0.043/0.026 = 1.67$ (Times).

The null hypothesis is

$$H_0 : RR = 1, H_1 : RR > 1.$$

The proportion of the amount of person-years for obesity group in the total person-years is $p = 699/2098 = 0.333$ and $1 - p = 0.667$. With $a = 30$ and $m = 66$, the probability $P(a \leq x \leq m)$ for random event x with $a \leq x \leq m$ in obesity group is obtained by using Eq. (25.10) as

$$P(30 \leq x \leq 66) = \sum_{x=30}^{66} \binom{66}{30} (0.333)^x (1 - 0.333)^{m-x} = 0.0270.$$

The probability is less than the given level of $\alpha = 0.05$. The null hypothesis H_0 is thus rejected. It is concluded that the relative risk of death in obesity women is significantly greater than that in non-obesity women.

By Eq. (25.11), the testing statistic is approximately equal to

$$Z = \frac{|30 - 66 \times 0.333| - 0.5}{\sqrt{66 \times 0.333 \times (1 - 0.333)}} = 1.960.$$

Referring to the table of standard normal distribution, the corresponding one-side probability is 0.0250. Here, the discrepancy between normal approximation and binomial probabilities for the data is only 0.002.

Relative risk is a comparative indicator. Its importance in public health varies with the disease frequency in control group. With the same level of relative risk, the higher the disease frequency in control group, the more importance the public health. More people could be protected by eliminating the exposure factor for the disease with high frequency in control group than for that with low frequency in control group.

25.3.2 *Attributable risk*

The attributable risk is defined as the difference of disease frequencies between exposed group and non-exposed group. It reflects the change of disease frequency caused by the exposure factor. It is sometimes called excess cumulative incidence probability or excess person-year incidence rate. With the same notations as above, the attributable risk in terms of cumulative incidence probability is estimated by

$$AR = q_1 - q_0 \quad (25.12a)$$

and the attributable risk in terms of person-year incidence rate is estimated by

$$AR = f_1 - f_0. \quad (25.12b)$$

Sometimes attributable risk may be expressed as a percentage and is called attributable risk fraction. It is estimated by

$$AR(\%) = \frac{q_1 - q_0}{q_1} \times 100\% \quad (25.13a)$$

and

$$AR(\%) = \frac{f_1 - f_0}{f_1} \times 100\%. \quad (25.13b)$$

For the data in Example 25.2 (Table 25.6) the attributable risk and attributable risk fraction are estimated by Eqs. (25.12a) and (25.13a) as

$$AR = 0.221 - 0.090 = 0.131$$

and

$$AR(\%) = \frac{0.221 - 0.090}{0.221} \times 100\% = 59.2\%.$$

For the data in Example 25.3 (Table 25.8) the attributable risk and attributable risk fraction are estimated by Eqs. (25.12b) and (25.13b) as

$$AR = 0.043 - 0.026 = 0.017$$

and

$$AR(\%) = \frac{0.043 - 0.026}{0.043} \times 100\% = 40.0\%.$$

25.3.3 Population attributable risk

The population attributable risk is defined as the difference of disease frequencies between the whole population and that in the control population. The estimate in terms of incidence probability is

$$PAR = q - q_0, \quad (25.14a)$$

where q is the incidence probability in the total population, q_0 is the incidence probability in control population. Population attributable risk can be expressed by percentage as

$$PAR(\%) = \frac{q - q_0}{q} \times 100\%, \quad (25.14b)$$

where $PAR(\%)$ is called the population attributable risk proportion.

q_0 and q in (25.14a) and (25.14b) can be substituted by person-year incidence rates f_0 and f . Thus we can obtain the corresponding population attributable risk and its proportion in terms of person-year incidence rate.

For the data in Example 25.2, if we use m_1/n as the estimate of incidence probability in total population, e.g., $q = 71/609 = 0.117$, by Eqs. (25.14a) and (25.14b), the population attributable risk and related proportion (%) are estimated as

$$PAR = 0.117 - 0.090 = 0.027,$$

$$PAR(\%) = \frac{0.117 - 0.090}{0.117} \times 100\% = 22.5\%.$$

The results show that comparing with the normal condition, the increase of the incidence probability of coronary heart disease in total population (27.0%) may be attributed to the high level of catecholamine in blood serum, or the proportion of 22.5% of CHD in the total population may be attributed to the high level of CAT. If the high level of CAT could be eliminated, the incidence of CHD in total population could have a net decrease (27.0%) or a relative decrease (22.5%) for the incidence probability.

For the data in Example 25.3, we use $f = 66/2098 = 0.031$ (or 31.0%) as person-year incidence rate in total population and $f_0 = 0.026$ as the substitutions for q and q_0 respectively, the population attributable risk and related proportion (%) are estimated as

$$PAR = 0.031 - 0.026 = 0.005 \text{ (or 5\%)}$$

$$PAR(\%) = \frac{0.031 - 0.026}{0.031} \times 100\% = 18.20\%.$$

25.3.4 Analysis of dose-response relationship

It is not unusual in practice that an etiological factor can be divided into several levels in terms of dosage from the lowest to the highest in order to uncover the dose-response relationship between disease and exposure. The layout of data is showed in Table 25.9 for this type of analysis. The column with subscript 0 is the information of non-exposed or lowest exposed group as control or reference group.

Table 25.9 can show if there is some trend of incidences with levels of exposure factor. However, to avoid the effect of sampling error on statistical inference, we have to work out a hypothesis test for trend.

H_0 : There is no linear trend between incidence and dose

H_1 : There is a linear trend between incidence and dose.

Table 25.9 The layout of data for dose-response analysis.

Exposure level	e_0	e_1	e_2	\cdots	e_k	
Score	x_0	x_1	x_2	\cdots	x_k	Total
Number of cases	a_0	a_1	a_2	\cdots	a_k	m_1
Number of non-cases	b_0	b_1	b_2	\cdots	b_k	m_0
Total number of subjects	n_0	n_1	n_2	\cdots	n_k	n
Incidence of disease	q_0	q_1	q_2	\cdots	q_k	

The statistic for trend test is

$$\chi^2 = \frac{\left(\sum_{i=0}^k a_i x_i - \frac{m_1}{n} \sum_{i=0}^k n_i x_i \right)^2}{\frac{m_1 m_0}{n^2(n-1)} \left[n \sum_{i=0}^k n_i x_i^2 - \left(\sum_{i=0}^k n_i x_i \right)^2 \right]}, \quad (25.15)$$

where x_i is the score for exposure level i . If the exposure levels are equally spaced, the values of x_i , $i = 1, 2, \dots$ can be changed as the integers in the natural order, 0, 1, 2, \dots and use the midpoints of the exposure levels as the scores. It can be proved, under the null hypothesis the statistic χ^2 follows a χ^2 distribution with one degree of freedom. Based on the value of χ^2 , the decision whether the null hypothesis H_0 should be rejected or not can be made. In essence, the process is to calculate the correlation coefficient between incidence frequency and dose and test the null hypothesis that the correlation coefficient equals to zero.

Equation (25.15) is also available for analyzing the relationship between person-year rate and exposure level.

Example 25.4 A prospective study was carried out to assess the relationship between a disease and sanitary situation. The data showed in Table 25.10 show the trend that the increase of disease incidence is accompanied by the decrease of sanitary situation. A trend test is worked out as follows.

Given $m_1 = 72$, $m_0 = 1326$, and $n = 1398$, from part two: calculation (at the bottom of Table 25.10) the sums are

$$\sum a_i x_i = 77, \quad \sum n_i x_i = 1175, \quad \sum n_i x_i^2 = 1761.$$

Table 25.10 Hypothesis testing for relationship between disease and sanitary situation.

Sanitary situation Score	e_i x_i	Good 0	Fair 1	Bad 2	Sum
Number of cases	a_i	19	29	24	72
Number of non-cases	b_i	497	560	269	1326
Total	n_i	516	589	293	1398
Incidence	q_i	0.037	0.049	0.082	
Relative risk	\hat{RR}_i	1.00	1.34	2.22	
Part two: calculation	$a_i x_i$	0	29	48	77
	$n_i x_i$	0	589	586	1175
	$n_i x_i^2$	0	589	1172	1761

By Eq. (25.15), the for testing trend is

$$\chi^2 = \frac{[77 - \frac{72}{1398} \times 1175]^2}{\frac{72 \times 1326}{1398^2(1398-1)} [1398 \times 1761 - 1175^2]} = 7.19. \quad (25.16)$$

Referring to the table of χ^2 distribution, it shows $P < 0.01$. The null hypothesis is rejected. There is significant linear relationship between disease and sanitary situation.

25.3.5 Stratified analysis

In some circumstances the statistical conclusion from overall analysis as above does not reflect the truth due to confounding effects. We will explain this with example.

Example 25.5 Subjects in an area were followed up for 9.5 years to explore the relationship between death from esophageal cancer and early symptoms in esophagus. The data are listed in Table 25.11. The overall or crude relative risk is $\hat{RR} = 2.717/1.616 = 1.68$.

By Eq. (25.11) for testing the null hypothesis $H_0 : RR = 1$, we have $Z = 4.11$, which is significant at $\alpha = 0.05$. Now we consider the possible confounding effect of age on the association between death of esophageal cancer and early symptoms in esophagus. The data are further stratified with age group showed in Table 25.12.

Table 25.11 Data from a prospective study on association between esophageal cancer and early symptoms in esophagus.

	Symptoms in esophagus		Total
	Yes	No	
Number of deaths from esophageal cancer	108	153	261
Amount of person-years	39756	94692	134448
Death rate per 1000 years	2.717	1.616	1.941

Table 25.12 Data stratified by age group for analyzing association between esophageal cancer and early symptoms in esophagus.

Age stratum	30-years			50-years			70-years		
	Yes	No	Total	Yes	No	Total	Yes	No	Total
No. of death	16	17	33	82	115	197	10	21	31
Person-year	22031	46149	68180	14814	39898	54712	2911	8645	11556
Death rate $^*(f_i)$	72.62	36.84		553.53	288.23		343.52	240.91	
\hat{RR}		1.97			1.92			1.43	

* Death rate per 100,000 person-years.

As comparison, the relative risks of the groups of 30-years and 50-years are greater than the crude relative risk $\hat{RR} = 1.68$; and that of 70-years is lower than the crude relative risk $\hat{RR} = 1.68$. This shows that there is some confounding by age group. In order to reflect the effect of early symptoms in esophagus, one may perform a stratified analysis. The procedure for stratified analysis is illustrated below.

25.3.5.1 Mantel-Haenzel stratified χ^2 test

The null hypothesis is H_0 : No association between esophageal cancer and early symptoms in esophagus in view of stratification; The alternative hypothesis is H_1 : there is some association between the cancer and early symptoms. Under H_0 , the test statistic

$$\chi^2 = \frac{\left[\sum a_i - \sum \left(\frac{m_{1i} T_{1i}}{T_i} \right) \right]^2}{\sum \frac{m_{1i} T_{1i} T_{0i}}{T_i^2}} \quad (25.17)$$

is distributed as $\chi^2_{(1)}$. The symbols a_i , m_{1i} denote the number of deaths and the total number of deaths from esophageal cancer in i th stratum; T_{1i} , T_{0i} and T_i denote the person-years for exposed group, non-exposed group, and both groups in i th stratum respectively; \sum denotes the summation over strata. The decision whether to reject the null hypothesis or not will be made according to the magnitude of χ^2 value.

With stratified χ^2 test, the results of Example 25.5 by Eq. (25.17) are below: The numerator is

$$\left[108 - \frac{33 \times 22031}{68180} - \frac{197 \times 14814}{54712} - \frac{31 \times 2911}{11556} \right]^2 = 1309.52.$$

And the denominator is

$$\begin{aligned} & \frac{33 \times 22031 \times 46149}{68180^2} + \frac{197 \times 14814 \times 39898}{54712^2} \\ & + \frac{31 \times 2911 \times 8645}{11556^2} = 51.96. \end{aligned}$$

Substituting these values into Eq. (25.17), we have

$$\chi^2 = \frac{1309.52}{51.96} = 25.20$$

the corresponding $P < 0.05$ such that H_0 is rejected. It concludes that there is significant linear relationship at the level of $\alpha = 0.05$ between esophageal cancer and early symptoms in esophagus.

The reasoning of Eq. (25.17) is explained as follows:

Under, H_0 the numerator in Eq. (25.17), $m_{1i}T_{1i}/T_i$, is the theoretical mean of a_i , and the denominator, $m_{1i}T_{1i}T_{0i}/T_i^2$, is the theoretical variance of a_i . Thus $\sum (m_{1i}T_{1i}/T_i)$ is the theoretical mean of $\sum a_i$, and $\sum (m_{1i}T_{1i}T_{0i}/T_i^2)$ is the theoretical variance of $\sum a_i$. Under the condition that H_0 is true, for a large sample, we have

$$\frac{\sum a_i - \sum \frac{m_{1i}T_{1i}}{T_i}}{\left[\sum \frac{m_{1i}T_{1i}T_{0i}}{T_i^2} \right]^{0.5}}$$

following $N(0, 1)$. Immediately, the square of the above fraction follows a χ^2 distribution with one degree of freedom.

25.3.5.2 Adjusted relative risk, RR_a

The Mantel-Haenszel method is used to estimate the adjusted relative risk for stratified data,

$$\hat{RR}_a = \frac{\sum \frac{a_i T_{0i}}{T_i}}{\sum \frac{b_i T_{1i}}{T_i}}, \quad (25.18)$$

where b_i is the number of deaths of non-exposed group in i th stratum.

By applying Eq. (25.18) to calculate the adjusted relative risk, the results for the data of Example 25.5 are as follows: The numerator is

$$\sum \frac{a_i T_{0i}}{T_i} = \frac{16 \times 46149}{68180} + \frac{82 \times 39898}{54712} + \frac{10 \times 8645}{11556} = 78.11$$

and the denominator is

$$\sum \frac{b_i T_{1i}}{T_i} = \frac{17 \times 22031}{68180} + \frac{115 \times 14814}{54712} + \frac{21 \times 2911}{11556} = 41.92.$$

Then the adjusted relative risk is

$$\hat{RR}_a = \frac{\sum \frac{a_i T_{0i}}{T_i}}{\sum \frac{b_i T_{1i}}{T_i}} = \frac{78.11}{41.92} = 1.86.$$

This value is different from the crude relative risk $\hat{RR} = 1.68$. This reflects that age group has a confounding effect on the relationship between esophageal cancer and early symptoms in esophagus.

The meaning of (25.18) is explained as follows: The estimate of relative risk in i th stratum is

$$\hat{RR}_i = \frac{\frac{a_i T_{0i}}{T_i}}{\frac{b_i T_{1i}}{T_i}}.$$

The adjusted relative risk is the pooled estimate over strata, that is, the ratio of the sum of numerators over strata divided by the sum of denominators over strata.

Table 25.13 Records of 12 subjects in a prospective study.

Subject <i>i</i>	Time entering into the study t_{in}	Time exiting from the study t_{out}	Result event
1	85-01-01	87-06-31(0)	1
2	85-01-01	88-06-31(0)	1
3	85-01-01	86-06-31(0)	0
4	86-01-01	88-06-31(0)	0
5	86-01-01	89-12-31	0
6	86-01-01	86-06-31(0)	1
7	87-01-01	87-06-31(0)	0
8	87-01-01	89-06-31(0)	0
9	87-01-01	89-12-31	0
10	88-01-01	89-12-31	0
11	88-01-01	89-06-31(0)	1
12	88-01-01	89-06-31(0)	0

25.3.5.3 Confidence interval for adjusted relative risk

The 95% confidence limits estimated by Eq. (25.9) are

$$1.86^{(1 \pm \frac{1.96}{\sqrt{25.20}})} = (1.46, 2.37).$$

25.4 Computerized Experiments

Experiment 25.1 Computation of interval cumulative incidences The SAS Program 25.1 is to compute the interval and cumulative incidences for data of 12 subjects in a prospective study showed in Table 25.13.

In Program 25.1, variable EVENT is an indicator of disease occurrence, coding 1 for case and 0 for non-case; WITHDRAW, FAIL and COUNT denote the number of withdrawals, the number of cases and the number of subjects in an interval respectively. TOTAL denotes the total number of subjects. Lines 01–23 are to create SAS data set INCID1 with the original data listed in Table 25.13 and to print. Date variables TIN and TOUT in lines 02 and 03 are defined as SAS date format, e.g., YYMMDD8. and YYMMDD9. FORMAT in line 04 is used to convert the format of YYMMDDw. to DATE7., where w. is the field length of date variable. Line 05 is to compute PERIOD, e.g. the length followed up. Lines 06–08 are to cut subject's PERIOD into INTERVALs. The corresponding output is showed

Program 25.1 Computing interval incidence and cumulative incidence.

Line	Program	Line	Program
01	DATA INCID1;		INTERVAL;
02	INPUT OBS TIN YYMMDD8. TOUT	28	DATA INCID2;
03	YYMMDD9. EVENT;	29	SET INCID1;
04	FORMAT TIN DATE7. TOUT DATE7.;	30	DROP OBS TIN TOUT PERIOD K;
05	PERIOD=(TOUT-TIN)/365; PUT PERIOD;	31	BY INTERVAL;
06	DO K=0 TO 4;	32	IF FIRST.INTERVAL THEN
07	IF K <=PERIOD<(K+1) THEN INTERVAL=K;	33	IF FIRST.INTERVAL THEN COUNT=0;
08	END;	34	COUNT+1;FAIL+EVENT;
09	CARDS;	35	TOTAL+1;
10	1 85-01-01 87-06-30 1	36	DROP EVENT;
11	2 85-01-01 88-06-30 1	37	IF LAST.INTERVAL THEN DO;
12	3 85-01-01 86-06-30 0	38	WITHDRAW=COUNT- FAIL;C=0;FREQ=FAIL;
13	4 86-01-01 88-06-30 0	39	OUTPUT;
14	5 86-01-01 89-12-31 0	40	WITHDRAW=COUNT- FAIL; C=1;FREQ=WITHDRAW;
15	6 86-01-01 86-06-30 1	41	OUTPUT;
16	7 87-01-01 87 06-30 0	42	END;
17	8 87-01-01 89-06-30 0	43	PROC PRINT DATA=INCID2;
18	9 87-01-01 89-12-31 0	44	TITLE 'INTERVAL';
19	10 88-01-01 89-12-31 0	45	DATA INCID3;
20	11 88-01-01 89-06-30 1	46	SET INCID2;
21	12 88-01-01 89-06-30 0	47	KEEP FAIL WITHDRAW FREQ INTERVAL C;
22	;	48	PROC LIFETEST DATA=INCID3
23	PROC PRINT;	49	INTERVALS=(0 TO 4) METHOD=ACT;
24	PROC SORT DATA=INCID1;	50	TIME INTERVAL*C(1);
25	BY INTERVAL;	51	FREQ FREQ;
26	PROC PRINT;	52	RUN;
27	FAIL=0;		
	TITLE 'SORTED DATA BY		

in Table 25.14. Lines 24–26 are to sort data INCID1 by (the number of) INTERVALs. The sorted results are showed in Table 25.15. Lines 28–42 are to resort data INCID1 for calculating the incidence probability of disease in the next step. The resorted dataset is named INCID2 and listed in Table 25.16. The code DROP in line 30 is to delete the variables which will be redundant in the later computation. Lines 31–42 are to sum up by intervals (line 33), to sum up the total (line 34) and to sum up the number

Table 25.14 SAS data set INCID1 created from the data in Table 25.1.

Subject OBS	Entering time T _{IN}	Exiting time T _{OUT}	Result EVENT	PERIOD	INTERVAL
1	01JAN85	30JUN87	1	2.49315	2
2	01JAN85	30JUN88	1	3.49589	3
3	01JAN85	30JUN86	0	1.49315	1
4	01JAN86	30JUN88	0	2.49589	2
5	01JAN86	31DEC89	0	4.00000	4
6	01JAN86	30JUN86	1	0.49315	0
7	01JAN87	30JUN87	0	0.49315	0
8	01JAN87	30JUN89	0	2.49589	2
9	01JAN87	31DEC89	0	3.00000	3
10	01JAN88	31DEC89	0	2.00000	2
11	01JAN88	30JUN89	1	1.49589	1
12	01JAN88	30JUN89	0	1.49589	1

Table 25.15 Sorted results of data INCID1.

OBS	T _{IN}	T _{OUT}	EVENT	PERIOD	INTERVAL
6	01JAN86	30JUN86	1	0.49315	0
7	01JAN87	30JUN87	0	0.49315	0
3	01JAN85	30JUN86	0	1.49315	1
11	01JAN88	30JUN89	1	1.49589	1
12	01JAN88	30JUN89	0	1.49589	1
1	01JAN85	30JUN87	1	2.49315	2
4	01JAN86	30JUN88	0	2.49589	2
8	01JAN87	30JUN89	0	2.49589	2
10	01JAN88	31DEC89	0	2.00000	2
2	01JAN85	30JUN88	1	3.49589	3
9	01JAN87	31DEC89	0	3.00000	3
5	01JAN86	31DEC89	0	4.00000	4

of cases (line 35). Lines 32 and 33 set $\text{FAIL} = 0$ and $\text{COUNT} = 0$ at the beginning of each interval (use FIRST.INTERVAL to identify). Lines 37–39 are to compute the number of withdrawals at the end of each interval (use LAST.INTERVAL to identify). A new variable C is used as an indicator of FREQ . $C = 0$ if FREQ represents FAILS or $C = 1$ if FREQ represents WITHDRAWS . Lines 39 and 41 are to print the new dataset. Variables C and FREQ are prepared for LIFETEST procedure. Line 43 is to print dataset INCID2 showed in Table 25.16. Lines 45–52 are to invoke

Table 25.16 Resorted data set INCID2.

OBS	INTERVAL	FAIL	COUNT	TOTAL	WITHDRAW	C	FREQ
1	0	1	2	2	1	0	1
2	0	1	2	2	1	1	1
3	1	1	3	5	2	0	1
4	1	1	3	5	2	1	2
5	2	1	4	9	3	0	1
6	2	1	4	9	3	1	3
7	3	1	2	11	1	0	1
8	3	1	2	11	1	1	1
9	4	0	1	12	1	0	0
10	4	0	1	12	1	1	1

Table 25.17 Data of 10 workers from a retro-prospective study.

Subject number	Gender sex	Birth date h4	Date entering h8	Date exiting s1	Disease occurrence h15
1	1	40-10-25	72-03-15	89-01-27	1
2	1	27-03-18	75-03-22	85-03-15	1
3	2	49-08-24	72-02-15	76-05-15	1
4	2	52-01-15	72-07-15	81-12-31	1
5	2	65-01-12	84-09-15	91-12-31	1
6	2	54-12-15	76-01-15	91-12-31	0
7	1	55-01-21	76-09-15	91-12-31	0
8	1	49-08-15	72-01-15	75-08-15	1
9	2	50-12-28	72-03-26	91-12-31	1
10	2	52-01-15	72-02-23	91-12-31	1

LIFETEST procedure. Since we choose METHOD = ACT, the life table method or actuarial method is required to compute cumulative incidence of disease development. Data set INCID3 with INTERVALS = (0 TO 4) is for the computation.

Experiment 25.2 Calculation of person-year incidence rates by age-time two way categories Data of 10 workers resulted from a retro-prospective study are showed in Table 25.17. (*h4*) denotes the birth date. (*h8*) denotes the date entering into the factory. (*s1*) is the date of disease occurrence with a dummy variable (*h15*) as an indicator of disease occurrence by coding $h15 = 1$ indicating disease and $h15 = 0$ indicating non-disease. In addition,

Program 25.2 Creating SAS permanent data set from data in Table 25.17.

Line	Program	Line	Program
01	LIBNAME AAA 'C:';	10	5 2 65-01-12 84-09-15 91-12-31 1
02	DATA AAA.JXSJ;	11	6 2 54-12-15 76-01-15 91-12-31 0
03	INPUT NUMBER SEX H4 YYMMDD8. H8 YYMMDD9. S1 YYMMDD9. H15;	12	7 1 55-01-21 76-09-15 91-12-31 0
04	FORMAT H4 DATE7. H8 DATE7. S1 DATE7.;	13	8 1 49-08-15 72-01-15 75-08-15 1
05	CARDS;	14	9 2 50-12-28 72-03-26 91-12-31 1
06	1 1 40-10-25 72-03-15 89-01-27 1	15	10 2 52-01-15 72-02-23 91-12-31 1
07	2 1 27-03-18 75-03-22 85-03-15 1	16	;
08	3 2 49-08-24 72-02-15 76-05-15 1	17	PROC PRINT DATA=AAA.JXSJ;
09	4 2 52-01-15 72-07-15 81-12-31 1	18	RUN;

the gender of workers are also recorded by coding *sex* = 1 if male and *sex* = 2 if female. (In the table, date is expressed as YY-MM-DD type.)

The computation is completed in the following three steps:

Step 1: Use Program 25.2 to create SAS permanent dataset named AAA.JXSJ. Date variables H4, H8, and S1 are input with the format yymmddw, where w is the width of entry. The system automatically converts the input date to SAS date. In line 04, the format date7. is defined to output the format dd-mm-yy.

Step 2: Run MACRO Program 25.3, which is composed by SAS Macro language, by click F10 key to create a MACRO statement in SAS system. The program begins with “%MACRO PRYEAR (MACRO variable 1, MACRO variable 2, ...)”, and ends with “%MEND”. PRYEAR is the name of the MACRO program, which closely follows %MACRO.

Variables in parentheses of line 01 are defined as follows: MINAGE=lower limit of age, MAXAGE=upper limit of age, SIZEAGE=interval of age group, MINYEAR=lower limit of time, MAXYEAR=upper limit of time, SIZEYEAR=time span of time group, EVENT = indicator of disease occurrence. These variables will be valued in the step 3.

Step 3: Use Program 25.4 to Invoke MACRO Program 25.3 for computing age-time grouped person-years and disease occurrences by gender.

Program 25.3 MACRO program for calculating person-years.

Line	Program
01	%MACRO PRYEAR(MINAGE=, MAXAGE=, SIZEAGE=, MINYEAR=, MAXYEAR=, SIZEYEAR=, EVENT=);
02	%IF &MINYEAR^= AND &MINAGE^= %THEN %DO;
03	%LET NAG=%EVAL((&MAXAGE-&MINAGE)/&SIZEAGE);
04	%LET NYR=%EVAL((&MAXYEAR-&MINYEAR)/&SIZEYEAR);
05	%DO I=1 %TO &NYR;
06	ARRAY PER&I.(J) PER&I.XP1-PER&I.XP&NAG;
07	%END;
08	%IF &EVENT NE %THEN %DO;
09	%DO I=1 %TO &NYR;
10	ARRAY &EVENT.PER&I.(J) PER&I.&EVENT.1-PER&I.&EVENT&NAG;
11	%END;
12	%END;
13	%END;
14	ARRAY TOTAL(P) PER1-PER&NYR;
15	%IF &EVENT NE %THEN %DO;
16	ARRAY T&EVENT.(P) &EVENT.PER1-&EVENT.PER&NYR;
17	%END;
18	NYIN=FLOOR((YIN-&MINYEAR)/&SIZEYEAR)+1;
19	NYDG=FLOOR((YDG-&MINYEAR)/&SIZEYEAR)+1;
20	NAGEIN=FLOOR((AGEIN-&MINAGE)/&SIZEAGE)+1;
21	NAGEDG=FLOOR((AGEDG-&MINAGE)/&SIZEAGE)+1;
22	DO J=NAGEIN TO NAGEDG;
23	DO P=NYIN TO NYDG;
24	TOTAL=MIN(AGEDG,&MINAGE+J*&SIZEAGE, (MDY(12,31,1900+
25	&MINYEAR+P*&SIZEYEAR)-BIRTHDT)/365.25)-MAX(AGEIN,
26	&MINAGE+(J-1)*&SIZEAGE (MDY(1,1,1900+&MINYEAR+(P-1)*
27	&SIZEYEAR)-BIRTHDT)/365.25);
28	IF TOTAL <=0 THEN TOTAL=0;
29	END;
30	END;
31	DO J=1 TO &NAG;
32	DO P=1 TO &NYR;
33	IF TOTAL <0 THEN TOTAL=0;
34	%IF &EVENT NE %THEN %DO;
35	IF (&MINAGE+(J-1)*&SIZEAGE <= AGEDG < &MINAGE+J*&SIZEAGE)
36	AND (&MINYEAR+(P-1)*&SIZEYEAR <= YDG < &MINYEAR+

(Continued)

Program 25.3 (Continued)

Line	Program
37	P*&SIZEYEAR) AND EVENT=1 THEN T&EVENT=1; ELSE T&EVENT=0;
38	%END;
39	END;
40	END;
41	%MEND;

Program 25.4 Invoking MACRO program to compute person-years and disease occurrences by age-time grouping.

Line	Program
01	LIBNAME AAA 'C:';
02	DATA AAA.PPY;
03	SET AAA.JXSJ;
04	BIRTHDT=H4;ENTRYDT=H8;EVENTDT=S1;
05	AGEIN=(ENTRYDT-BIRTHDT)/365;AGEDG=(EVENTDT-BIRTHDT)/365;
06	YIN=YEAR(ENTRYDT)-1900;YDG=YEAR(EVENTDT)-1900;
07	%PRYEAR(MINAGE=10,MAXAGE=90,SIZEAGE=10,
08	MINYEAR=60,MAXYEAR=95,SIZEYEAR=5,EVENT=H15)
09	PROC MEANS SUM;
10	CLASS SEX;
11	RUN;

Lines 02–06 in Program 25.4 are used to create new SAS permanent data set AAA.PPY from AAA.JXSJ by renewing variable names. AGEIN is the age entering the factory. AGEDG is the age of disease occurrence ($S1=1$) or the age withdrawal from observation ($S1=0$). YIN and YDG are entering year and disease developing year respectively with 1900 as start point. Line 07 invokes MACRO program %MACRO PRYEAR. The closely followed variable names in parentheses should be kept consistent with the line 01 in MACRO program 25.3 in order to assign values to these variables. For value assignment it is required that all intervals of age groups are equal and all spans of time groups are equal. Moreover, the upper limits, lower limits and differences between upper and lower limits are integer times of the interval or span. If only values are assigned to MINAGE=, MAXAGE=, and SIZEAGE=, then the MACRTO

can only compute the person-years and disease occurrences by age grouping of one-way classification, and not by time grouping. In the same way, if only values are assigned to variables MINYEAR=, MAXYEAR= and SIZEYEAR=, the MACRTO can only compute the person-years and disease occurrences by time grouping, and not by age grouping. Invoked by Program 25.4, the MACRO program %MACRO PRYEAR yields 131 variables. Among them there are 56 variables, which record the person-years of every subject contributed to each combination of age-time grouping, named PER1XP1, PER1XP2, ..., PER1XP8, ..., PER7XP1, ..., PER7XP7, PER7XP8. The numbers as suffix of PER indicate the year group in 5-year span, and the numbers as suffix of XP indicate the age group in 10 years interval. Accordingly, the 56 variables named PER1H151, ..., PER1H158, PER2H151, ... and PER7H158 record the number of disease occurrences by age-time combination. Lines 09 and 10 take sum for each variable by gender.

The amount of person-years and the numbers of disease occurrences for the data in Table 25.17 are computed and shown in Table 25.18. Based on these values, the person-year incidence rates can be obtained easily.

Table 25.18 Person-years and disease occurrences by age-time grouping.

Time grouping (5-year span)		Age grouping (10-year interval)				
		Age1 (10–)	Age2 (20–)	Age3 (30–)	Age4 (40–)	Age5 (50–60)
yr1	(1970–)	0	3.029 (11.296)	2.859	0	0
yr2	(1975–)	0	3.998 (20.661①)	5.081	1.956	2.873
yr3	(1980–)	0 (0.312)	5.055 (10.053①)	0.845 (7.270)	4.268	5.084
yr4	(1985–)	0 (0.030)	0.053 (5.051)	5.028 (15.244)	4.104①	0.240①
yr5	(1990–95)	0	0 (2.014①)	2.021 (5.033①)	(1.036①)	

Note: (1) In the cells, the upper values are for male and the lower values in parentheses are for female. (2) The symbol ① placed at the end of values indicates 1 case of disease occurrence.

Table 25.19 Leukemia cases among patients suffering from polycythemia vera and receiving different radiotherapy.

Treatment group	Number of patients	Number of leukemia cases
Non treatment	133	1
X-ray treatment	79	7
P32 treatment	228	25
X-ray + P32 treatment	72	12
Total	512	45

Table 25.20 Data from a historical prospective study on relationship between asbestos and death from lung cancer.

		Workers exposed to asbestos		Workers non-exposed		
Years worked	Person-years	Deaths	Rate per	Person-years	Deaths	Rate per
		from lung cancer	10,000 person-years		from lung cancer	10,000 person-years
10–	89462	36	4.02	74395	14	1.88
20–	51925	164	31.58	62528	86	13.75
30–	17001	177	104.11	19360	96	45.59
40–	8465	109	128.77	7236	41	56.66
Total	166853	486	29.13	163519	237	14.49

25.5 Practice and Experiments

1. Table 25.19 shows the data from a follow-up study on leukemia cases among patients suffering from polycythemia vera and treated with different radiotherapy. Analyze incidences probabilities and relative risks of leukemia in different treatment groups.
2. Table 25.20 shows the data from a historical prospective study on relationship between lung cancer and asbestos. Try to explore the effect of asbestos on death from lung cancer by using stratified method, and compare the adjusted relative risk with crude relative risk.
3. Analyze the data in Table 25.20 again for exposure group to explore whether there is an increased trend of death rate from lung cancer with the increase of age.

(1st edn. and 2nd edn. Songlin Yu, Jiqian Fang)

Chapter 26

Designs and Analysis of Case-Control Studies

Case-control study is well known as a retrospective study. Based on the occurrence of disease the method makes inference about the possible factors that cause the disease. The study follows a paradigm that proceeds backward from disease to exposure, e.g., “from result looking back to cause”. Two kinds of individuals are needed in the study, those with the disease of interest called “cases” and those without the disease called “controls”. The ratio of exposure proportion in the past among cases to that among controls can offer some evidence for the association between disease and exposure. The association provides clues for further pragmatic study. It is widely used in etiological research for chronic diseases and investigation for causes of disease outbreak.

26.1 Designs of Case-Control Studies

26.1.1 *Types of designs*

There are two types of designs in case-control studies.

26.1.1.1 *Design for group comparison*

A group of patients who have a specific disease is selected as the case group, and a group of people who do not have the disease serves as the control group. The histories of individuals in the two groups are compared. For instance, Goldsmith *et al.* selected 88 diagnosed patients of bladder cancer as a case group and 258 healthy people as a control group, and the

authors found that the level exposed to hydride of heavy metal and organic accelerator was higher in the case group than in the control group. This type of design is simple and easy to carry out. Its disadvantage is that the result is influenced easily by confounding factors and thus leads to instability. When more variables are to be considered, methods of stratified analysis or multivariate regression models may be applied to control the effects from confounding factors.

26.1.1.2 *Design for matched comparison*

In order to eliminate the effects of confounding factors on investigated results, each case is individually matched to a set of controls (usually one or two, but sometimes more), which have similar values as the cases do for several important confounding variables such as gender, age, race, occupation, personal or family history of disease and so on. The analysis is based on such matched sets to promote statistical power. When one case matches one control only, it is called 1:1 matched design or paired design. When one case matches two controls, it is called 1:2 matched design. One case can match as many as four controls, but if the number of controls in each set exceeds 4, the statistical power increases slightly. As data from matched design cannot provide any information about the variables used for match, attention as to be paid when selecting matched variables. Non-confounding factors should not be selected as matched variables in order to avoid the so-called over matching.

26.1.1.3 *Case-crossover design*

The case-crossover design was introduced in 1991 by M. Maclure (*Am. J. Epidemiol.* (1991) 133, 144–153) to study effects of transient short-term exposure on the risk of acute events, including rare acute-onset disease. The design involves exposure levels of cases and the exposure ones of their own when they do not fall in any case. Each case serves as his/her own control. In comparison of exposure level of event onset with the level of event absence, we can obtain the information about the difference between the two levels associated with different event onset situation.

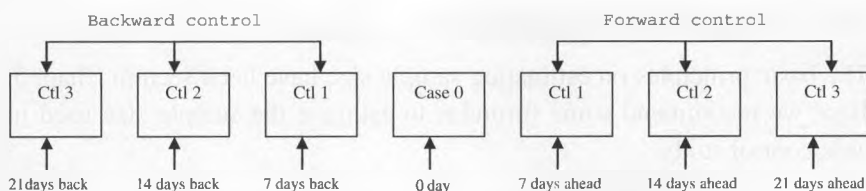


Fig. 26.1 A 1:6 bi-directional case-crossover design with control period of 7 days for each.

This design bears resemblance to both a classic crossover study and a matched case-control study. For the former each subject serves as his/her own control. For the latter, the inference is based on a comparison of exposure distribution rather than the disease.

In case-crossover study, cases serve as their own controls and therefore the design eliminates confounding by stable individual characteristics. The potential bias may come from two sources in time series study. The first is the trend and seasonality. Because the design is to compare exposure levels at different time points, trend or seasonality can confound the comparison. The second is the effect overlap. If the time span selected for comparison between case and control is too short, the control may still stay in the hazard period, so that the exposure effect may be under estimated.

Several approaches may be available for selecting the control period. Unidirectional selection uses time period before the event occurs; and the bi-directional selection uses both time periods before and after the event occurs. In each direction selection 1:1 (1 case : 1 control) matching, or 1: M (1 case : M ($M > 1$) controls) be used. As an example, if we conduct a study on the relationship between air pollution and death, the case-crossover design applies a 1:3 bi-directional approach and control period equals 7 days. Each death serves as case and backward 7, 14 and 21 days as retrospective controls, as well as forward 7, 14 and 21 days as prospective controls. The mode can be found in Fig. 26.1.

It is showed from Fig. 26.1 that with bi-directional 1:6 matching design and control period of 7 days, each case matches 3 control periods in either direction. The total time span is $21 \times 2 + 1 = 43$ days, 21 days backward and 21 days forward, and 1 day belongs to case itself.

26.1.2 Determination of sample size

The basic principles on estimating sample size have been seen in Chap. 5. Here we recommend some formulas to estimate the sample size used in case control study.

26.1.2.1 Sample size for group comparison

Estimation is completed in two steps. The first step is to calculate N' by using the formula

$$N' = \frac{[Z_\alpha \sqrt{(1+C)PQ} + Z_\beta \sqrt{P_1Q_1 + P_0Q_0C}]^2}{(P_1 - P_0)^2}, \quad (26.1)$$

where C = ratio of the number of cases to the number of controls is given by investigator in advance. For example, if equal sample sizes in both case and control groups are planned then $C = 1.0$; If the number of individuals in case group is half that in control group, then $C = 0.50$.

P_0 is the estimated proportion of individuals exposed to risk factor in the control population. $Q_0 = 1 - P_0$ is the proportion of individuals without exposure in the control population;

$$P_1 = \frac{P_0 RR}{[1 + P_0(RR - 1)]}$$

is the estimated proportion of cases exposed to risk factor in the case population. $Q_1 = 1 - P_1$ is the proportion of cases without exposure in the case population; RR is the estimate of relative risk under the alternative hypothesis; and

$$P = \frac{P_1 + P_0}{2}, \quad Q = 1 - P.$$

Let Z_α be the standard normal deviation with the probability of type I error α . It is usual to give $\alpha = 0.05$, thus the one-side value is $Z_{0.05} = 1.645$, and the two-side value is $Z_{0.05} = 1.96$. Z_β is the standard normal deviation with the probability of type II error β . It is usual to give one-side $\beta = 0.10$, thus $Z_{0.1} = 1.282$.

The second step is to calculate sample size N for case group. The formula is

$$N = \left(\frac{N'}{4} \right) \left(1 + \sqrt{1 + \frac{4}{N'\delta}} \right)^2, \quad (26.2)$$

where $\delta = |P_1 - P_0|$.

Example 26.1 A case-control study is planned to explore the association between chronic obstructive disease of lungs and smoking. The proportion of smokers in the control population is about 48%. The null hypothesis is $H_0: RR = 1.0$, i.e., there is no association between chronic obstructive disease of lungs and smoking. The alternative hypothesis is $H_1: RR = 3.0$, i.e., there is some positive association between the disease and smoking. Let $\alpha = 0.05$, $\beta = 0.10$, and $C = 1.0$, the process to estimate the sample size is illustrated as follows:

Solution $P_0 = 0.48$, $Q_0 = 0.52$, $RR = 3.0$, $C = 1.0$, $Z_{0.05} = 1.645$ and $Z_{0.1} = 1.282$, we have

$$P_1 = \frac{0.48 \times 3.0}{1 + 0.48(3.0 - 1)} = 0.7347,$$

$$Q_1 = 1 - 0.7347 = 0.2653,$$

$$P = (0.5)(0.7347 + 0.48) = 0.6073,$$

$$Q = 1 - 0.6073 = 0.3927.$$

According to Eq. (26.1) we have

$$\begin{aligned} N' &= \frac{[1.645\sqrt{(1+1)(0.6073)(0.3927)} \\ &\quad + 1.282\sqrt{(0.7347)(0.2653) + (0.48)(0.52)}]^2}{(0.7347 - 0.48)^2} \\ &= 61.09 \end{aligned}$$

N' and $\delta = 0.7347 - 0.48 = 0.2547$ are substituted into Eq. (26.2) and the sample size N is

$$N = \frac{61.09}{4} \left[1 + \sqrt{1 + \frac{4}{61.09 \times 0.2547}} \right]^2 = 68.72 \approx 69.$$

That is, 69 cases and 69 controls are required for the study.

If we choose $C = 0.50$, i.e., the number of subjects in case group is half that in control group, from Eq. (26.1) we obtain

$$N' = \frac{[1.645\sqrt{(1+0.50)(0.6073)(0.3927)} + 1.282\sqrt{(0.7347)(0.2653) + (0.48)(0.52)(0.50)}]^2}{(0.7347 - 0.48)^2} = 45.01$$

N' and δ are substituted into Eq. (26.2), the number of cases is

$$N = \frac{45.01}{4} \left[1 + \frac{4}{45.01 \times 0.2547} \right]^2 = 52.57 \approx 53.$$

Therefore, 53 cases and $53 \times 2 = 106$ controls are required.

26.1.2.2 Sample size required for matched comparison

Let N represent the number of matched sets. For 1:1 matched design the formula for estimating N is

$$N \approx \frac{M}{P_0 Q_1 + P_1 Q_0}, \quad (26.3)$$

where P_0 is the proportion of exposed individuals in control population, and the corresponding non exposed proportion is $Q_0 = 1 - P_0$. P_1 is the proportion of exposed cases in case population, and the corresponding non-exposed proportion is $Q_1 = 1 - P_1$. The calculation of P_1 is explained in Eq. (26.1). M , the number of matched sets in which the case and the control are discordant in exposure, is calculated by

$$M = \frac{[(0.50)Z_\alpha + Z_\beta \sqrt{P(1-P)}]^2}{(P - 0.50)^2}, \quad (26.4)$$

where P is estimated by

$$P \approx \frac{RR}{1 + RR}. \quad (26.5)$$

When $m > 1$, i.e., 1 : m matching, at first we have N by Eq. (26.3), then the adjusted number of matched sets N' is estimated by

$$N' = \frac{N(1+m)}{2m}. \quad (26.6)$$

We again use Example 26.1 of a case-control study on exploration of the association between smoking and chronic obstructive disease of lungs to illustrate the calculation of sample size for matched design. If 1:1 matched design is adopted, the procedure of sample size needed is calculated as follows:

$$P = \frac{3.0}{1 + 3.0} = 0.75,$$

$$M = \frac{[(0.50)1.645 + 1.282\sqrt{(0.75)(0.25)}]^2}{(0.75 - 0.5)^2} = 22.042.$$

By Eq. (26.3), the number of cases as well as controls needed is

$$N = \frac{22.042}{(0.48)(0.2653) + (0.7347)(0.52)} = 43.27 \approx 44.$$

If 1:2 matched design is adopted, the number of cases needed is

$$N' = \frac{44(1 + 2)}{2 \times 2} = 33.$$

Since each case is matched with two controls, 66 controls are needed.

26.1.3 Selection of cases and controls

Before proceeding a case-control study, consideration must be given to the diagnostic criteria for defining the disease of interest and the eligibility criteria for selection of cases and controls in order to avoid selecting individuals without the disease as cases or selecting individuals with latent or untypical disease of interest as controls. There are two main approaches of cases and controls. One is based on hospitalized patients; patients with the disease concerned are grouped into the case group, and some of the patients without the given disease are grouped into the control group. Another approach is based on local population; all new patients, or a random sample from them, collected from disease registry or medical network within a period of time serve as cases; controls are selected randomly from disease-free individuals in the population. The latter strategy is available only for the areas with integrated medical services system and disease registration or report system. Otherwise it is difficult to collect all new cases.

26.1.4 Bias and Avoidance

Since case-control study is retrospective, it is possible that in the whole process of the study various biases may be introduced into the study and finally the reliability of the ultimate results can be influenced. Often the situations, where biases may possibly occur, are

1. When resource is based on hospitalization, the admission rate of patients with the same symptoms may have different disease. Therefore, spurious association of disease and the characteristics may occur. This bias is called Berkson fallacy.
2. With ambiguous diagnostic criteria, patients without that disease might be wrongly taken into the case group, or reverse, patients with the latent, light, or untypical disease might be wrongly taken into the control group.
3. In the phase of information collection, bias might be introduced into the data because of the differences in time, in place, in mode of observers, or dim memory of subjects when the field investigation is carried out.

In implementation of a case-control study one should deliberate the source of subjects, rigorously hold eligibility criteria of cases and controls, provide unified investigation method and the order of questioning for interview. If it is possible the blinding method could be applied, which leads to field investigators knowing nothing about who is case and who is control; in such a way some effects from investigator's subjective factors can be eliminated.

26.2 Analysis of Data from Design for Group Comparison

Data from case-control studies cannot be used to calculate the incidence probability or incidence rate and relative risk, which should be based on the data from prospective studies. The odds ratio (*OR*), sometimes denoted with ψ , is an essential indicator used to reflect the difference of exposure between cases and controls, and to establish the association between disease of interest and the exposure. In this section we will discuss the methods used to analyze the data from the design for group comparison. In the next section

Table 26.1 2×2 table (data layout of dichotomized exposure for case-control study).

Group	Exposure		Total
	Presence	Absence	
Case	a	b	n_{1+}
Control	c	d	n_{0+}
Total	n_{+1}	n_{+0}	n

we will discuss the methods used to analyze the data from the design for matched comparison.

Analysis of data for group comparison usually begins with a simple 2×2 table related to single factor; then succeeds to stratified methods related to multiple factors, and dose-response relationship, etc.

26.2.1 Analysis of data for a single 2×2 table

For a case-control study if the exposed history can be dichotomized as present/absent or high/low among cases and controls, and the subjects can be regarded homogeneous in other aspects, then the data can be organized as a single 2×2 table showed in Table 26.1.

26.2.1.1 Calculation of odds ratio

As discussed before, the odds is the ratio of the probability of an event occurrence to the probability of an event non-occurrence. We use the ratio of exposed proportion to non-exposed proportion in each group as an estimate of odds for the event of exposure. For example, the estimate of exposed odds in case group is

$$odd_1 = \frac{\frac{a}{n_{1+}}}{\frac{b}{n_{1+}}} = \frac{a}{b}.$$

The estimate of exposed odds in control group is

$$odd_0 = \frac{\frac{c}{n_{0+}}}{\frac{d}{n_{0+}}} = \frac{c}{d}.$$

The ratio of the exposed odds for case group to the exposed odds for control group is defined as the estimated odds ratio, denoted by $\hat{\psi}$. We have

$$\hat{\psi} = \frac{odd_1}{odd_0} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}. \quad (26.7)$$

It has been proved that the ratio of two odds of exposure for different disease status from case-control study equals the ratio of two odds of disease occurrence for different exposures from prospective study. In addition, when the probability of disease occurrence is low (for example, less than 1%), the odds ratio approximates to the relative risk. Both of these two points are the essential theoretical basis why case-control study can be widely used for etiological research and why odds ratio is so important in medical statistics.

Example 26.2 In a case-control study on relationship between cardiac infarction and use of oral contraceptive drug, 234 cases with cardiac infarction and 1746 controls were investigated about their current use of oral contraceptive drug. Data are showed in Table 26.2. Analyze them.

Solution The odds ratio is estimated by Eq. (26.7) as

$$\hat{\psi} = \frac{ad}{bc} = \frac{29 \times 1607}{205 \times 135} = 1.68.$$

It is showed that the odds of current use of oral contraceptive drug in the case group is 1.68 times as much as that in the control group.

Table 26.2 Data from case-control study on relationship between cardiac infarction and current use of oral contraceptive drug.

Group	Current use of oral contraceptive drug		Total
	Yes	No	
Case	29	205	234
Control	135	1607	1742
Total	164	1812	1976

26.2.1.2 Hypothesis testing for odds ratio

The odds ratio estimated from sample $\hat{\psi}$ exists sampling variation. Statistical inference for relationship between disease and exposure can be made only after consideration of sampling error. The hypothesis test for odds ratio is

$$H_0 : \psi = 1.0, \quad H_1 : \psi \neq 1.0.$$

The test statistic is the same as Mantel-Haenszel χ^2 statistic in Chap. 25. That is

$$\chi^2 = \frac{(n-1)(ad-bc)^2}{n_{1+}n_{0+}n_{+1}n_{+0}}. \quad (26.8)$$

Under the null hypothesis, this χ^2 statistic follows χ^2 distribution with one degree of freedom. For Example 26.2, the χ^2 statistic is

$$\chi^2 = \frac{(1976-1)(29 \times 1607 - 205 \times 135)^2}{164 \times 1812 \times 234 \times 1742} = 5.84$$

the value is greater than 3.84, the upper side critical value of χ^2 distribution with one degree of freedom given $\alpha = 0.05$. The null hypothesis is thus rejected at the level of $\alpha = 0.05$. The result shows that the odds ratio is significantly different from 1.0, there exists significant relationship between cardiac infarction and current use of oral contraceptive drug.

26.2.1.3 Confidence interval for odds ratio

Exact estimation of the confidence interval for odds ratio is rather complicated. We discuss here two simple methods to estimate the confidence interval approximately, which are accurate enough for most practical situation.

- (1) Woolf's method As $\hat{\psi}$ is ranged between $(0, \infty)$, by logarithmic transformation, $\ln \hat{\psi}$ is ranged between $(-\infty, \infty)$ and approximately follows a normal distribution, of which the mean is 0 and the variance is given by

$$\text{Var}(\ln \hat{\psi}) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}. \quad (26.9)$$

The 95% confidence limits for $\ln \psi$ are approximated as

$$\ln \hat{\psi} \pm 1.96\sqrt{\text{Var}(\ln \hat{\psi})} \quad (26.10a)$$

thus the 95% confidence limits for ψ are estimated as

$$\hat{\psi} \exp\left(\pm 1.96\sqrt{\text{Var}(\ln \hat{\psi})}\right). \quad (26.10b)$$

For Example 26.2, the variance of $\ln \hat{\psi}$ is

$$\text{Var}(\ln \hat{\psi}) = \frac{1}{29} + \frac{1}{205} + \frac{1}{135} + \frac{1}{1607} = 0.0474.$$

Substituting 0.0474 into Eq. (26.10b), we obtain the estimates of 95% confidence limits for ψ as

$$(1.68) \exp(\pm 1.96\sqrt{0.0474}) \text{ or } (1.10, 2.57).$$

- (2) Miettinen's method Substitute the χ^2 statistic directly calculated by Eq. (26.8) into the following equation to yield the 95% confidence limits for

$$\hat{\psi}^{\left(1 \pm \frac{1.96}{\sqrt{\chi^2}}\right)}. \quad (26.11)$$

For Example 26.2, the odds ratio is $\hat{\psi} = 1.68$ and the test statistic is $\chi^2 = 5.84$. Substituting these values into Eq. (26.11), result in the approximate 95% confidence limits

$$(1.68)^{\left(1 \pm \frac{1.96}{\sqrt{5.84}}\right)} \text{ or } (1.10, 2.56).$$

26.2.1.4 Population attributable risk

Calculations of population attributable risk are given in Eqs. (25.14a) and (25.14b) of Chap. 25 for data from prospective design. Let D denotes dichotomous disease variable (with D and \bar{D} representing presence and absence of the disease respectively) and a dichotomous exposure variable E (with E and \bar{E} representing exposed and unexposed levels respectively).

It is not difficult to prove that

$$\begin{aligned}
 PAR &= \frac{P - P_0}{P} = \frac{P(D) - P(D|\bar{E})}{P(D)} = 1 - \frac{P(D|\bar{E})}{P(D)} \\
 &= 1 - \frac{P(D|\bar{E})}{P(D|E)P(E) + P(D|\bar{E})P(\bar{E})} = \frac{1}{(RR)P(E) + P(\bar{E})} \\
 &= 1 - \frac{1}{P(E)(RR - 1) + 1}. \quad (26.12)
 \end{aligned}$$

If the control group is a random sample selected from the general population, then we use the exposed proportion in the control group c/n_{0+} as a substitution for the exposed probability $P(E)$ in population, and use $\hat{\psi}$ as a substitution for relative risk RR . Thus, the equation for calculating population attributable risk in case-control study is

$$PAR = 1 - \left[\frac{c}{n_{0+}} (\hat{\psi} - 1) + 1 \right]^{-1} = 1 - \frac{bn_{0+}}{dn_{1+}}. \quad (26.13)$$

With the data listed in Table 26.2 for Example 26.2, we have $b = 205$, $n_{0+} = 1742$, $d = 1607$ and $n_{1+} = 234$. By Eq. (26.13) the estimated population attributable risk is

$$PAR = 1 - \frac{bn_{0+}}{dn_{1+}} = 1 - \frac{205 \times 1742}{1607 \times 234} = 0.050.$$

It shows that 5% of the new cases of cardiac infarction in the population is attributable to the recent use of oral contraceptive drug.

26.2.2 Analysis for stratified 2×2 tables

When data from case-control study are stratified by possible confounding factors, under the condition that the exposure is a dichotomous variable, the whole dataset can be divided into k strata, and there is a 2×2 table for each strata. The layout of 2×2 table in stratum i ($i = 1, 2, \dots, k$) is showed in Table 26.3.

Example 26.3 For the data in Example 26.2, the odds ratio was calculated as $\hat{\psi} = 1.68$ from a crude 2×2 table shown in Table 26.2. To eliminate the possible confounding effect of age on the odds ratio, the data are stratified

Table 26.3 Data layout of 2×2 table in stratum i .

Group	Exposure		Total
	Yes	No	
Case	a_i	b_i	n_{1i}
Control	c_i	d_i	n_{0i}
Total	m_{1i}	m_{0i}	n_i

by age group into five subsets of 2×2 tables showed in the first part of Table 26.4. Calculate the 5 odds ratios and see what happen.

Solution Most $\hat{\psi}_i$ are greater than the crude odds ratio $\hat{\psi} = 1.68$ except that of age group 35–39. The results show that age has some confounding effect on the association between cardiac infarction and current use of oral contraceptive drug. It leads to the crude odds ratio much lower than stratified odds ratios.

A comprehensive analysis for data in stratified 2×2 tables includes estimating adjusted odds ratio, hypothesis testing and confidence interval. Two widely used methods are introduced below.

26.2.2.1 *M–H adjustment*

The method is suggested by N. Mantel and W. Haenszel (1959). Adjusted odds ratio ψ_{M-H} is estimated by

$$\hat{\psi} = \frac{\sum_{i=1}^k \left(\frac{a_i d_i}{n_i} \right)}{\sum_{i=1}^k \left(\frac{b_i c_i}{n_i} \right)}. \quad (26.14)$$

The M–H testing statistic for stratified analysis has the form

$$\chi_{M-H}^2 = \frac{\left(\sum_{i=1}^k \frac{a_i d_i - b_i c_i}{n_i} \right)^2}{\sum_{i=1}^k \left(\frac{n_{1i} n_{0i} m_{1i} m_{0i}}{(n_i - 1) n_i^2} \right)}. \quad (26.15)$$

Under H_0 , the statistic χ_{M-H}^2 follows χ^2 distribution with one degree of freedom. The decision to whether reject H_0 or not is made based on the magnitude of the value of χ_{M-H}^2 .

The 95% confidence interval for odds ratio is

$$\hat{\psi} \left(1 \pm \frac{1.96}{\sqrt{\chi^2_{M-H}}} \right) \quad (26.16)$$

Data in Example 26.3 are further divided by age groups and showed in Table 26.4 for M-H stratified analysis. The lower part of the table lists the quantities ready for stratified analysis.

The adjusted odds ratio is

$$\hat{\psi} = \frac{23.7060}{5.9714} = 3.97.$$

The value is greater than the crude odds ratio ($\hat{\psi} = 1.68$). It shows that after eliminating the confounding effects of age, the association between cardiac infarction and current use of oral contraceptive drug appears more obviously.

For testing the null hypothesis $H_0 : \psi = 1.0$, by Eq. (26.15) the χ^2_{M-H} statistic is

$$\chi^2_{M-H} = \frac{17.7346^2}{9.2641} = 33.95.$$

It is greater than the critical value $\chi^2_{0.05(1)} = 3.84$. The null hypothesis is therefore rejected. The conclusion is that there is significant association between cardiac infarction and current use of oral contraceptive drug.

Using Eq. (26.16), the 95% confidence limits for the adjusted odds ratio ψ are

$$3.97^{(1 \pm \frac{1.96}{\sqrt{33.95}})} \text{ or } (2.50, 6.31).$$

26.2.2.2 Logarithmic transformation adjustment

The equation for calculating adjusted odds ratio is

$$\hat{\psi} = \exp \left(\frac{\sum_{i=1}^k W_i \ln \hat{\psi}_i}{\sum_{i=1}^k W_i} \right), \quad (26.17)$$

where W_i , the weight for stratum i , is the reciprocal of variance for stratum i . The formulas for W_i is

$$W_i = \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right)^{-1}. \quad (26.18)$$

Table 26.4 Data of cardiac infarction and current use of oral contraceptive drug (stratified by age group).

Current use of oral contraceptive drug	Age group (year)															Total
	25-29			30-34			35-39			40-44			45-49			
	<i>d</i>	<i>c</i>	Total	<i>d</i>	<i>c</i>	Total	<i>d</i>	<i>c</i>	Total	<i>d</i>	<i>c</i>	Total	<i>d</i>	<i>c</i>	Total	
Users	4	62	66	9	33	42	4	26	30	6	9	15	6	5	11	
Non-users	2	224	226	12	390	402	33	330	363	65	362	427	93	301	394	
Total	6	286	292	21	423	444	37	356	393	71	371	442	99	306	405	
(1) $\hat{\psi}_i$		7.23			8.86			1.54			3.71			3.88		
(2) $a_i d_i / n_i$		3.0685			7.9054			3.3588			4.9140			4.4593		23.7060
(3) $b_i c_i / n_i$		0.4247			0.8919			2.1832			1.3235			1.1481		5.9714
(4) $(a_i d_i - b_i c_i) / n_i$		2.6438			7.0135			1.1756			3.5905			3.3112		17.7346
(5) $\frac{n_{1i} n_{0i} m_{1i} m_{0i}}{(n_i - 1) n_i^2}$		1.0316			1.7174			2.3692			2.1646			1.9813		9.2641
(6) $\ln \hat{\psi}_i$		1.9782			2.1815			0.4318			1.3110			1.3558		
(7) w_i		1.2977			4.3992			3.1076			3.3792			2.6265		14.8102
(8) $w_i \ln \hat{\psi}_i$		2.5671			9.5972			1.3418			4.4302			3.5615		21.4978

Note: *d* denotes the number of cases, *c* the number of controls.

The 95% confidence limits for adjusted odds ratio ψ is

$$\hat{\psi} \exp \left(\pm \frac{1.96}{\sqrt{\sum_{i=1}^k W_i}} \right). \quad (26.19)$$

As illustration, the logarithmic transformation method is applied to the data in Example 26.3. Quantities prepared for calculating $\hat{\psi}$ are listed in rows (6), (7) and (8) of Table 26.4. By formula (26.17) the adjusted odds ratio is

$$\hat{\psi} = \exp \left(\frac{21.4978}{14.8102} \right) = 4.27.$$

The 95% confidence limits for ψ are

$$4.27 \exp \left(\pm \frac{1.96}{\sqrt{14.8102}} \right) \quad \text{or} \quad (2.57, 7.11).$$

Notice that the confidence interval here is similar to the result of M-H adjustment.

26.2.3 Analysis of dose-response for several exposure categories

When exposure has several ordinal categories, it is possible to test if dose-response relationship appears, i.e., the higher (or lower) the exposure level, the greater (or less) the odds ratio. The null hypothesis is

H_0 : There is no dose-response relationship

H_1 : A linear dose-response relationship exists.

The test statistic is

$$\chi^2 = \frac{\left(\sum_{i=0}^k a_i x_i - \frac{n_1}{n} \sum_{i=0}^k m_i x_i \right)^2}{\frac{n_1 n_0}{n^2 (n-1)} \left[n \sum_{i=0}^k m_i x_i^2 - \left(\sum_{i=0}^k m_i x_i \right)^2 \right]}, \quad (26.20)$$

where k is the number of ordinal exposure categories, x_i is the ordered value for each category. Under H_0 , the statistic χ^2 follows χ^2 distribution with one degree of freedom. Based on the value of χ^2 , a decision whether the null hypothesis H_0 should be rejected or cannot be made.

Table 26.5 Data from a case-control study on assessing association between pyreticosis and amount eaten of raw cotton seed oil.

Category of amount eaten (kg/year)	≤ 6	7-8	9-10	≥ 11	
Ordered value	0	1	2	3	Total
case: a_i	15	53	44	24	136 (n_1)
control: b_i	63	168	104	26	361 (n_0)
Total number of subjects: m_i	78	221	148	50	497 (n)
odd: s_i	0.2381	0.3155	0.4231	0.9231	
Odds ratio: $\hat{\psi}_i$	1.0	1.30	1.75	3.82	
Calculation: $a_i x_i$	0	53	88	72	213 ($\sum a_i x_i$)
$m_i x_i$	0	221	296	150	667 ($\sum m_i x_i$)
$m_i x_i^2$	0	221	592	450	1263 ($\sum m_i x_i^2$)

Example 26.4 The data from a case-control study on assessing the association between pyreticosis and the intake amount of cotton seed oil are listed in Table 26.5. Judge whether there is a linear dose-response relationship.

Using Eq. (26.20) the χ^2 statistic is

$$\chi^2 = \frac{[213 - \frac{136}{497}(667)]^2}{\frac{136 \times 361}{(497)^2 \times (497-1)}[497 \times 1263 - (667)^2]} = 12.68.$$

The value of $\chi^2 = 12.68$ is greater than $\chi^2_{0.05(1)} = 3.84$. We conclude that there is a significant dose-response relationship between pyreticosis and the amount of cotton seed oil taken.

26.3 Analysis of Matched Data

26.3.1 Analysis of data with 1:1 matching

When one case is matched to one control and the exposure under study is dichotomous, n matched sets can be categorized to one of the four possible exposure combinations showed in Table 26.6.

The display of four combinations of exposure and disease status in Table 26.6 can be reorganized to the 1:1 matched data layout shown in Table 26.7 for easy understanding.

Table 26.6 Four possible combinations of dichotomous exposure for 1:1 matched data.

	Exposure combination				Total
	(1)	(2)	(3)	(4)	
Case	+	+	—	—	
Control	+	—	+	—	
Number of matched sets	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>n</i>

Note: + denotes exposed, — denotes non-exposed.

Table 26.7 Data layout for 1:1 matched design.

Case's exposure status	Control's exposure status		Total
	+	—	
+	<i>a</i>	<i>b</i>	<i>a + b</i>
—	<i>c</i>	<i>d</i>	<i>c + d</i>
Total	<i>a + c</i>	<i>b + d</i>	<i>n</i>

It is showed from the two tables above that the terms *a* and *d* denote the numbers of sets in which both of the case and control were accordantly exposed (++) or non-exposed (--) to the study factor respectively. Those accordant sets can bring no information to the relationship of disease and exposure. While the terms *b* and *c* denote the numbers of sets in which only the case or only the control was exposed and symbolized as (+-) or (-+). Those discordant sets bring on information to the relationship between disease and exposure.

The estimate of the odds ratio, conditional on the matched design, is

$$\hat{\psi} = \frac{b}{c}. \quad (26.21)$$

As with the χ^2 square test for matched data in Chap. 6, for the hypothesis test

$$H_0 : \psi = 1, \quad H_1 : \psi \neq 1$$

Table 26.8 Data of 303 matched sets from a 1:1 matched case-control study on relationship of leukemia and occupational exposure to benzene.

Case's exposure history to benzene	Control's exposure history to benzene		Total
	Yes	No	
Yes	2	30	32
No	12	259	271
Total	14	289	303

the statistic used is

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}, \quad (26.22)$$

Under the null hypothesis, the statistic follows χ^2 distribution with one degree of freedom. According to the value of χ^2 statistic, a decision can be made whether the null hypothesis should be rejected or not.

The 95% confidence interval of ψ is estimated by

$$\hat{\psi} \left(1 \pm \frac{1.96}{\sqrt{\chi^2}} \right). \quad (26.23)$$

Example 26.5 The data of 303 matched sets showed in Table 26.8 were resulted from a 1:1 matched case-control study on the relationship of leukemia and occupational exposure to benzene.

By Eqs. (26.21)–(26.23), the odds ratio and test statistic are

$$\begin{aligned} \hat{\psi} &= \frac{30}{12} = 2.5, \\ \chi^2 &= \frac{(|30 - 12| - 1)^2}{30 + 12} = 6.88, \end{aligned}$$

$P < 0.05$ so that there is a significant association between leukemia and the history of occupational exposure to benzene. The 95% confidence limits for ψ are

$$2.5 \left(1 \pm \frac{1.96}{\sqrt{6.88}} \right) \text{ or } (1.26, 4.96).$$

Table 26.9 Possible outcomes for matched set i ($i = 1, 2, \dots, k$) in 1:m matched design.

	Exposed	Unexposed	Total
Case	Y_i	$1 - Y_i$	1
Control	X_i	$m - X_i$	m

26.3.2 Analysis of data with 1:m matching

The layout of data for the design of 1 case matched by m controls appears complicated. Let k denotes the number of matched sets, the possible outcomes in i th matched set ($i = 1, 2, \dots, k$) are displayed in Table 26.9.

In Table 26.9, Y_i is an indicator, $Y_i = 1$ if the case is exposed, and $Y_i = 0$ if not. X_i represents the number of exposed, and $(m - X_i)$ the number of non-exposed among m controls. k matched sets are regarded as k strata. By using Eq. (26.14) the odds ratio is

$$\hat{\psi} = \frac{\sum_{i=1}^k Y_i(m - X_i)}{\sum_{i=1}^k X_i(1 - Y_i)}. \quad (26.24)$$

For test

$$H_0 : \psi = 1, \quad H_1 : \psi \neq 1$$

as Eq. (26.15), the test statistic is calculated by

$$\chi^2_{M-H} = \frac{(\sum mY_i - \sum X_i)^2}{\sum (1 + m)(X_i + Y_i) - \sum (X_i + Y_i)^2}. \quad (26.25)$$

Example 26.6 In order to explore the possible association of development of infectious hepatitis with taking meals in public restaurants within 40 days before the disease occurrence, a 1:4 matched design was adopted. One case was matched with four controls in gender, age group, and no hepatitis patient in control's family. The data of 18 matched sets are listed on the left of Table 26.10. Work out the analysis.

Table 26.10 Data from a 1:4 case-control study on the association of infectious hepatitis with meals in public restaurants.

No. of set i	Part one: Raw data			Part two: Quantity for calculating odds ratio							
	Exp. of case Y_i	Exp. of control X_i	No. of control m_i	$m_i Y_i$	$1 - Y_i$	$X_i(1 - Y_i)$	$m_i - X_i$	$Y_i(m_i - X_i)$	$X_i + Y_i$	$(1 + m_i)$ $(X_i + Y_i)$	$(X_i + Y_i)^2$
1	0	0	4	0	1	0	4	0	0	0	0
2	1	1	4	4	0	0	3	3	2	10	4
3	1	0	4	4	0	0	4	4	1	5	1
4	0	0	4	0	1	0	4	0	0	0	0
5	0	1	4	0	1	1	3	0	1	5	1
6	1	0	4	4	0	0	4	4	1	5	1
7	1	0	4	4	0	0	4	4	1	5	1
8	0	0	4	0	1	0	4	0	0	0	0
9	1	2	4	4	0	0	2	2	3	15	9
10	1	1	4	4	0	0	3	3	2	10	4
11	1	2	4	4	0	0	2	2	3	15	9
12	0	0	4	0	1	0	4	0	0	0	0
13	1	4	4	4	0	0	0	0	5	25	25
14	1	1	4	4	0	0	3	3	2	10	4
15	1	1	4	4	0	0	3	3	2	10	4
16	1	1	4	4	0	0	3	3	2	10	4
17	0	0	4	0	1	0	4	0	0	0	0
18	1	2	4	4	0	0	2	2	3	15	9
Total	12	16	72	48	6	1	56	33	28	140	76

Note: $Y_i = 1$ if case is an exposed, $Y_i = 0$ if case is a non-exposed. X_i denotes the exposed number among controls in the i th set.

Solution The basic quantities prepared for further use are listed on the right of Table 26.10. By Eq. (26.25) the value of the testing statistic is

$$\chi^2_{M-H} = \frac{(48 - 16)^2}{140 - 76} = 16.0.$$

It is greater than $\chi^2_{0.01(1)} = 6.63$. Then $H_0 : \psi = 1$ is reject at $\alpha = 0.01$, and the conclusion is that the development of infectious hepatitis significantly associates with taking meals in public restaurants within 40 days before the disease occurs. Using Eq. (26.24), the estimate of odds ratio is

$$\hat{\psi} = \frac{\sum_{i=1}^k Y_i(m - X_i)}{\sum_{i=1}^k X_i(1 - Y_i)} = \frac{33}{1} = 33.$$

By Eq. (26.23), the 95% confidence limits for ψ are

$$33 \left(1 \pm \frac{1.96}{\sqrt{16.0}} \right) \text{ or } (5.95, 183).$$

26.3.3 Analysis of data with 1 : m_i matching

In matched design, it is usual that 1 case matches a fixed number, m , of controls for all sets in the study. But it is not unusual that 1 or more controls may be lost in some sets. This results in unequal number of controls. Let m_i denotes the number of controls in i th set. The sets with the equal number of controls can be put together for analysis.

Example 26.7 Suppose that in Example 26.6 we have additional data which consist of five sets with 1:1 matching and eight sets with 1:3 matching. The supplementary data are listed in Table 26.11, where $m_i = 1$ or 3. We assemble the two parts of data in Tables 26.10 and 26.11 as a whole to calculate the odds ratio.

Having combined two parts of data, by using Eq. (26.25),

$$\chi^2_{M-H} = \frac{(48 + 21 - 16 - 9)^2}{140 + 60 - 76 - 34} = 21.51.$$

It is greater than $\chi^2_{0.05(1)} = 3.84$, $P < 0.05$. We reject $H_0 : \psi = 1$ at $\alpha = 0.05$. It shows that there is a significant association between developing infectious hepatitis and taking meals in public restaurants within 40 days before occurrence of the disease.

Table 26.11 Supplementary data to Example 26.6 shown in Table 26.10.

No. of set i	Part one: Raw data			Part two: Quantity for calculating odds ratio							
	Exp. of case Y_i	Exp. of control X_i	No. of control m_i	$m_i Y_i$	$1 - Y_i$	$X_i(1 - Y_i)$	$m_i - X_i$	$Y_i(m_i - X_i)$	$X_i + Y_i$	$(1 + m_i)$ $(X_i + Y_i)$	$(X_i + Y_i)^2$
19	1	0	1	1	0	0	1	1	1	2	1
20	1	0	1	1	0	0	1	1	1	2	1
21	1	1	1	1	0	0	0	0	2	4	4
22	0	1	1	0	1	1	0	0	1	2	1
23	0	1	1	0	1	1	0	0	1	2	1
24	1	1	3	3	0	0	2	2	2	8	4
25	1	1	3	3	0	0	2	2	2	8	4
26	1	1	3	3	0	0	2	2	2	8	4
27	1	1	3	3	0	0	2	2	2	8	4
28	1	2	3	3	0	0	1	1	3	12	9
29	1	0	3	3	0	0	3	3	1	4	1
30	0	0	3	0	1	0	3	0	0	0	0
31	0	0	3	0	1	0	3	0	0	0	0
Total	9	9	29	21	4	2	20	14	18	60	34

Program 26.1 Computation of adjusted odds ratio for data stratified 2 × 2 tables shown in Table 26.4.

Line	Program	Line	Program
01	DATA STRATIFY;	09	4 62 2 224 9 33 12 390 4 26 33 330
02	DO I=1 TO 5; /* age group */	10	6 9 65 362 6 5 93 301
03	DO A=1 TO 2;	11	;
04	DO B=1 TO 2;	12	PROC FREQ;
05	INPUT F @@;	13	WEIGHT F;
06	OUTPUT;	14	TABLES A*B/ALL NOCOL;
07	END; END; END;	15	TABLES I*A*B/ALL NOCOL;
08	CARDS;	16	RUN;

Using Eq. (26.24), the odds ratio is estimated as

$$\hat{\psi} = \frac{33 + 14}{1 + 2} = 15.67.$$

Using Eq. (26.23), the estimated 95% confidence limits for $\hat{\psi}$ are

$$15.67^{(1 \pm \frac{1.96}{\sqrt{21.51}})} \text{ or } (4.90, 50.13).$$

26.4 Computerized Experiments

Experiment 26.1 Computation of odds ratio for stratified 2 × 2 Tables

Program 26.1 computes odds ratio adjusted for age with the data in Table 26.4, which are stratified to 5 2 × 2 Tables.

In Program 26.1, lines 09 and 10 are the data listed in Table 26.4. Line 14 computes the crude odds ratio. Line 15 computes the age-stratified odds ratios and age-adjusted odds ratio. Table 26.12 lists part of the output from the program.

Experiment 26.2 Computation of odds ratio for data from 1:1 matched design

In SAS software there is no special procedure to compute odds ratio and corresponding variance for matched data. PROC logistic can be used to compute the odds ratio instead. The odds ratio for data in Example 26.5 can be obtained by SAS Program 26.2.

Variables X_1 and X_2 in line 2 of Program 26.2 denote the exposed levels of case and control respectively, with 1 if the subject is exposed and 0 if not. Variable F denotes the cell frequency. Line 03 creates a constant

Table 26.12 Output from program 26.1 for data listed in Table 26.4.

Age-stratification i	Odds ratio $\hat{\psi}_i$	95% confidence interval	Mantel-Haenszel χ^2	P -value
1	7.226	1.293–40.374	6.776	0.009
2	8.864	3.482–22.565	28.642	<0.0001
3	1.538	0.506–4.667	0.583	0.445
4	3.713	1.278–10.783	6.583	0.010
5	3.884	1.159–13.016	5.533	0.019
Age-adjusted	3.970	2.510–6.280	34.723	<0.0001
Crude	1.684	1.104–2.570	5.841	0.016

Program 26.2 Computation of odds ratio and corresponding standard error for data of Example 26.5 with 1:1 matched design.

Line	Program	Line	Program
01	DATA MATCHING;	06	;
02	INPUT X1 X2 F @@;	07	PROC LOGISTIC;
03	Y = 1; X = X1 - X2; OUTPUT;	08	MODEL Y = X/NOINT;
04	CARDS;	09	WEIGHT F;
05	1 1 2 1 0 30 0 1 12 0 0 259	10	RUN;

Table 26.13 Criteria for assessing model fit.

Criterion	Without covariates	With covariates	Chi-square for covariates
$-2 \log L$	420.047	412.077	7.970 with 1 DF ($P = 0.0048$)
Score	—	—	7.714 with 1 DF ($P = 0.0055$)

response variable Y to satisfy mechanical operation. Variable X in the same line is used to compare the odds of (1,0) and (0,1). Results are showed in Tables 26.13 and 26.14.

In Table 26.13, $-2 \log L = 420.047 - 412.077 = 7.970$ is the likelihood ratio χ^2 statistic with $P = 0.0048$; Score = 7.714 is the value of χ^2 statistic for score test with $P = 0.0055$. The two P -values are closed to each other. These test statistics show that the odds of exposure in case is significantly different from that in control. Parameter estimate is 0.9163 (see Table 26.14),

Table 26.14 Analysis of maximum likelihood estimates.

Variable	Parameter estimate	Standard error	Wald χ^2	Pr > χ^2	Standardized estimate
X	0.9163	0.3416	7.1965	0.0073	1.865984

Table 26.15 Data of bladder cancer and smoking history.

Group	Smoking history		Total
	Exposed	unexposed	
Cases with bladder cancer	192	129	321
Controls	156	181	337
Total	348	310	658

and its standard error is 0.3416, odds ratio is $\exp(0.9163) = 2.50$. That means the odds of exposure in cases is 2.5 times as large as that in controls. The 95% confidence limits are $\exp(0.9163 \pm 0.3146)$ or (1.280, 4.883).

26.5 Practice and Experiments

1. Table 26.15 listed below is the sorted outcome from a case-control study on association between bladder cancer and smoking history. Analyze the data.
2. Table 26.16 is the sorted outcome from a case-control study on association between myocardiac infarction onset and alcohol consumption pre-onset. Compute the odds ratio and test for trend.
3. Derive Eqs. (26.24) and (26.25) in Sec. 26.3 from Eqs. (26.14) and (26.15) in Sec. 26.2.
4. In Sec. 26.2, we set two propositions. The first one is that in case-control study, the ratio of two odds of exposures in different disease status equals the ratio of two odds of disease occurrences in different exposures. And the second is that when the probability of disease occurrence is low (for example, less than 1%), the odds ratio approximates the relative risk.
 - (1) Let D and \bar{D} denote case and control respectively, E and \bar{E} denote exposure and non-exposure respectively. Use the conditional

Table 26.16 Distribution of amount of alcohol consumption of 391 myocardial infarction Patients and 418 controls.

Amount of alcohol consumption per day (g)	Patient suffering from myocardial infarction	Control
0	136	110
0-100	202	238
100-200	42	46
250+	11	24
Total	391	418

Table 26.17 Data from a 1:1 matched case-control study on relationship of breast cancer and breast-feeding history.

Breast-feeding history for control	Breast-feeding history for patient with breast cancer		Total
	Yes	No	
Yes	27	65	92
No	23	43	66
Total	50	108	158

probabilities $P(E | D)$, $P(\bar{E} | D)$, $P(E | \bar{D})$, $P(\bar{E} | \bar{D})$, $P(D | E)$, $P(\bar{D} | E)$, $P(D | \bar{E})$ and $P(\bar{D} | \bar{E})$ to express the two propositions.

- (2) Use Bayse' formula to prove the equivalence between the above mentioned two types of odds ratios.
- (3) Prove that under the condition of low disease incidence, the odds ratio approximates the relative risk.
5. Table 26.17 is the sorted outcome from a 1:1 matched case-control study on relationship between breast cancer and breast-feeding history. Compute the odds ratio to explore the effects of breast feeding history on breast cancer occurrence. If the data would be treated as if they were come from a design for group comparison, what value would be the odds ratio?

(1st edn. and 2nd edn. Songlin Yu)

Chapter 27

Design and Analysis of Diagnostic and Screening Tests

The main purpose of clinical diagnosis and epidemiologic screening is to detect at early stage if a person has a related disease. Some medical studies therefore aim at finding out or creating better strategies for diagnostic or screening test. Problems arising in this kind of studies include: How to design? How to analyze the data? How to evaluate the usability of a diagnostic test or screening test? How to compare the difference among diagnostic tests or screening tests? We will deal with these problems in this chapter.

27.1 Design and Data Layout

Subjects should be those with or without a certain disease confirmed by gold standard, which is the evidence obtained by more refined methods such as results from biopsy, surgical operation, pathological anatomy or autopsy, X-ray film, CT scan, long-term following up, or other convincing tests. Case and control subjects were recruited, respectively, from the population with the disease and without the disease. The sample sizes for each group should be larger than 20 (Youden, 1950). Test results can be expressed as Table 27.1.

27.2 Measures Frequently Used in Diagnostic Test

Example 27.1 113 prostate cancer patients were diagnosed by radioimmunoassay of prostate acid phospholipase (RIA-PAP), and 217 patients

Table 27.1 Result of a diagnostic test.

True disease status	Test results		Total
	Positive (T_+)	Negative (T_-)	
Case (D_+)	A	B	$A + B$
Control (D_-)	C	D	$C + D$
Total	$A + C$	$B + D$	N

Table 27.2 RIA-PAP diagnoses prostate cancer.

True disease status	RIA-PAP results		Total
	+	-	
Cancer	79	34	113
No cancer	13	204	217
Total	92	238	330

without prostate cancer were taken as controls. Table 27.2 shows the diagnostic test results. The prevalence of prostate cancer in community is 35/100000. Compute relevant measures of diagnostic test to assess the results in the table.

The following are the measures frequently used in evaluation of diagnostic test:

27.2.1 Sensitivity

The definition of sensitivity is the probability of positive testing result if the disease is present, also called true positive rate (TPR), and denoted as Sen .

$$Sen = P(T_+ | D_+) = A/(A + B) = TPR. \quad (27.1)$$

The standard error is

$$SE_{Sen} = \sqrt{AB/(A + B)^3} = \sqrt{Sen(1 - Sen)/(A + B)}.$$

For Example 27.1, $Sen = 79/113 = 0.6991$, i.e., true positive rate $TPR = 0.6991$. Among the patients with prostate cancer, 69.91% are positive; its

standard error is

$$SE_{Sen} = \sqrt{0.6991(1 - 0.6991)/113} = 0.0431 = 4.31\%.$$

27.2.2 Specificity

Specificity is defined as the probability of negative testing result if the disease is absent.

$$Spe = P(T_- | D_-) = D/(C + D). \quad (27.2)$$

The standard error is

$$SE_{Spe} = \sqrt{CD/(C + D)^3} = \sqrt{Spe(1 - Spe)/(C + D)}.$$

For the example, $Spe = 204/217 = 0.9401$, i.e., among the patients without prostate cancer, 94.01% are negative; its standard error is

$$SE_{Spe} = \sqrt{0.9401(1 - 0.9401)/217} = 0.0161 = 1.61\%.$$

From Table 27.1 we can get the probability of a false negative $\beta = 1 - Sen = B/(A + B)$ and the probability of a false positive $\alpha = 1 - Spe = C/(C + D)$. α is also called false positive rate (*FPR*).

For Example 27.1, the probability of a false negative is estimated as $\beta = 1 - Sen = 1 - 0.6991 = 0.3009$; the probability of a false positive, or false positive rate is $\alpha = 1 - Spe = 1 - 0.9401 = 0.0599$, or $FPR = 0.0599$. The relationship between sensitivity and specificity is displayed in Fig. 27.1, where the cross point between the middle vertical line and the horizontal axis is called the cutoff point. Using the critical point, subjects can be classified as positive or negative.

The value of sensitivity or specificity ranges from 0 to 1. The closer to 1 the value, the better is the accuracy of a test. When we compare two diagnostic tests, and use the two measures sensitivity and specificity only, it may occur that one test has higher sensitivity and lower specificity and the other has lower sensitivity and higher specificity. In this case, we could not determine which one is better. Therefore, other measures that combine sensitivity and specificity, such as Youden's index, positive likelihood ratio, and negative likelihood ratio, are suggested for evaluating a diagnostic test.

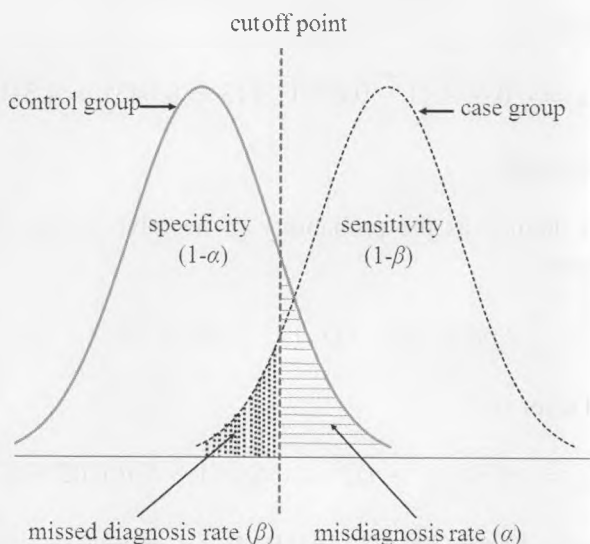


Fig. 27.1 The relationship between sensitivity and specificity.

27.2.3 Youden's index

Youden's index is the difference between true positive rate (TPR) and false positive rate (FPR).

$$J = TPR - FPR = Sen - (1 - Spe). \quad (27.3)$$

The standard error is

$$\begin{aligned} SE_J &= \sqrt{AB/(A+B)^3 + CD/(C+D)^3} \\ &= \sqrt{Sen(1-Sen)/(A+B) + Spe(1-Spe)/(C+D)}. \end{aligned}$$

The value of Youden's index ranges from 0 to 1. The closer to 1 the value, the better is the accuracy of a test.

In Example 27.1, $J = 0.6991 - 0.0599 = 0.6392$, i.e., Youden's index is 0.6392; its standard error is

$$SE_J = \sqrt{0.6991(1 - 0.6991)/113 + 0.9401(1 - 0.9401)/217} = 0.0461.$$

27.2.4 Positive likelihood ratio

Positive likelihood ratio is the ratio of true positive rate (TPR) to false positive rate (FPR).

$$LR_+ = \frac{TPR}{FPR} = \frac{Sen}{1 - Spe}. \quad (27.4)$$

In Example 27.1, $LR_+ = 0.6991/0.0599 = 11.67$.

The range of LR_+ values is from 0 to ∞ . The larger the value is, the stronger is the ability of a test to confirm a disease.

27.2.5 Negative likelihood ratio

Negative likelihood ratio is the ratio of false negative rate to true negative rate, i.e., the ratio of missed diagnosis rate to specificity.

$$LR_- = \frac{1 - TPR}{1 - FPR} = \frac{1 - Sen}{Spe}. \quad (27.5)$$

In Example 27.1, $LR_- = (1 - 0.6991)/(1 - 0.0599) = 0.32$.

The values of LR_- also range from 0 to ∞ . The smaller the value is, the stronger is the ability of a test to confirm a disease.

27.2.6 Positive predictive value

Positive predictive value is the probability that a person actually has the disease when the test is positive.

$$\begin{aligned} PV_+ = P(D_+ | T_+) &= \frac{P(D_+)P(T_+ | D_+)}{P(D_+)P(T_+ | D_+) + P(D_-)P(T_+ | D_-)} \\ &= \frac{P_0 Sen}{P_0 Sen + (1 - P_0)(1 - Spe)}, \end{aligned} \quad (27.6)$$

where $P_0 = P(D_+)$ denotes the population prevalence,

$$P(D_-) = 1 - P(D_+) = 1 - P_0.$$

For Example 27.1, $P_0 = 0.00035$, $Sen = 0.6991$, $Spe = 0.9401$, substitute them into Eq. (27.6), we have

$$\begin{aligned} PV_+ = P(D_+ | T_+) &= \frac{0.00035 \times 0.6991}{0.00035 \times 0.6991 + (1 - 0.00035)(1 - 0.9401)} \\ &= 0.0041 \approx 1/246. \end{aligned}$$

If the whole population is diagnosed by RIA-PAP, a patient, whose test result is positive actually has prostate cancer among 246 patients. It reflects the worthiness of the prediction when the test result is positive.

27.2.7 Negative predictive value

Negative predictive value is the probability that a person actually is disease-free given that the test is negative.

$$\begin{aligned} PV_- &= P(D_- | T_-) \\ &= \frac{P(D_-)P(T_- | D_-)}{P(D_-)P(T_- | D_-) + P(D_+)P(T_- | D_+)} \\ &= \frac{(1 - P_0)Spe}{(1 - P_0)Spe + P_0(1 - Sen)}. \end{aligned} \quad (27.7)$$

Substitute $P_0 = 0.00035$, $Sen = 0.6991$, $Spe = 0.9401$ into Eq. (27.7), we have

$$\begin{aligned} PV_- &= P(D_- | T_-) \\ &= \frac{(1 - 0.00035) \times 0.9401}{(1 - 0.00035) \times 0.9401 + 0.00035 \times (1 - 0.6991)} \\ &= 0.9999. \end{aligned}$$

Among 10,000 patients whose test results are negative, 9999 patients actually do not have prostate cancer, i.e., only 1 patient has the disease. It is taken as an indication of the worthiness of the prediction when the test result is negative.

Note that only when the sample prevalence $P_1 = (A + B)/N$ is equal to the population prevalence $P_0 = P(D_+)$, then Eq. (27.6) can be reduced to

$$PV_+ = \frac{A}{A + C} \quad (27.8)$$

and Eq. (27.7) can be reduced to

$$PV_- = \frac{D}{B + D} \quad (27.9)$$

In Example 27.1, according to Eqs. (27.8) and (27.9), we can get $PV_+ = 79/92 = 0.86$, $PV_- = 204/238 = 0.86$. They are very different from the results of Eqs. (27.6) and (27.7). This is because the sample prevalence $P_1 = (A + B)/N = 113/330 = 0.34$ is not equal to the population prevalence $P_0 = 0.00035$. In this case, to compute the predictive value, one should use Eqs. (27.6) and (27.7), and not Eqs. (27.8) and (27.9).

In other words, if the case and control groups are a random sample from the population, rather than two samples from disease and disease-free subpopulation respectively, we can substitute the sample prevalence $P_1 = (A + B)/N$ for population prevalence. In such a case, Eqs. (27.8) and (27.9) are suitable. Otherwise, Eqs. (27.6) and (27.7) must be used.

The values of PV_+ and PV_- range between 0 and 1. When the prevalence of the population is fixed, the closer to 1 the value is, the better is the accuracy of a test.

27.3 Analysis of ROC Curve

It must be noted that the values of true positive rate (TPR) and false positive rate (FPR) depend on the cutoff point for a diagnostic test based on continuous variable. To evaluate comprehensively the accuracy of a diagnostic test, all possible cutoff points should be considered.

ROC curve is the abbreviation for "receiver operating characteristic curve" or "relative operating characteristic curve", originated from telecommunications for evaluating the quality of receiving signal. Since the '80s, *ROC* curve analysis has been widely applied to evaluate the performance of a diagnostic test in medicine. Through changing the diagnostic cutoff points, pairs of TPR and FPR can be obtained. Taking FPR as x -axis, TPR as y -axis, a *ROC* curve can be obtained from plotting TPR against FPR for all possible cutoff points. The area under the curve denoted as AUC is a new measure to assess the usability of a diagnostic test.

Table 27.3 Hypothetical data of a continuous variable.

Gold standard		Test results				
Case group	15.90	13.35	12.87	10.22	5.01	
Control group	8.29	6.24	4.61	1.77		

27.3.1 Calculating ROC operating point

The variables for *ROC* curve analysis can simply be classified into two types, continuous variable and ordinal variable. Roughly speaking the variables measured by laboratory tests are continuous, while those of clinical image diagnosis and psychological evaluation are ordinal.

Example 27.2 A study on diagnostic test has five patients in the case group and four patients in the control group. The test results are shown in Table 27.3. Calculate all possible *TPR* and *FPR* (note that the sample sizes are too small).

Solution Rank the nine values from large to small and use the first eight values (do not include the smallest value 1.77) as diagnostic cutoff point. The test result is positive if the value is larger than or equal to the cutoff point, otherwise negative. The results are shown in by the following eight fourfold tables:

cutoff point = 15.90			cutoff point = 13.35			cutoff point = 12.87			cutoff point = 10.22		
Test			Test			Test			Test		
Group	+	−	Group	+	−	Group	+	−	Group	+	−
Case	1	4	Case	2	3	Case	3	2	Case	4	1
Control	0	4	Control	0	4	Control	0	4	Control	0	4

cutoff point = 8.29			cutoff point = 6.24			cutoff point = 5.01			cutoff point = 4.61		
Test			Test			Test			Test		
Group	+	−	Group	+	−	Group	+	−	Group	+	−
Case	4	1	Case	4	1	Case	5	0	Case	5	0
Control	1	3	Control	2	2	Control	2	2	Control	3	1

Table 27.4 *FPR* and *TPR* for the data in Table 27.3.

	Diagnostic cutoff point							
	15.90	13.35	12.87	10.22	8.29	6.24	5.01	4.61
<i>FPR</i>	0	0	0	0	1/4	2/4	2/4	3/4
<i>TPR</i>	1/5	2/5	3/5	4/5	4/5	4/5	5/5	5/5

Table 27.5 Diagnostic categories of 109 CT films.

Gold standard	CT results					Total
	1	2	3	4	5	
Abnormal	3	2	2	11	33	51
Normal	33	6	6	11	2	58

For each of the above fourfold tables, we can calculate a pair of (*FPR*, *TPR*) which is called operating points (or coordinate points) of a *ROC* (see Table 27.4). If there are same values (ties) among the test results, we only need to keep one value among the tie as the cutoff point.

Example 27.3 Among 109 films of CT, 51 films were diagnosed as abnormal, 58 films were diagnosed as normal by a gold standard. A radiologist classified the whole set of films into grade 1, 2, 3, 4 and 5 according to his confidence degree to abnormal. The results are shown in Table 27.5. Calculate all possible (*FPR*, *TPR*).

Solution Rank the results from large to small and use the first four categories (do not include the smallest value 1) as diagnostic cutoff points respectively. The CT result is positive if a category value is larger than or equal to the cutoff point, otherwise, negative. Similar to Example 27.2, the results can also be expressed in fourfold tables. From all fourfold tables with different operating points, *FPR* and *TPR* can be calculated which are shown in Table 27.6.

Table 27.6 it is assumed that larger test result indicates more positive. If smaller test result indicates more positive, we should rank the test results from small to large and assume it is positive if a value is smaller than or equal to the cutoff point, otherwise, negative.

Table 27.6 *FPR* and *TPR* for the data in Table 27.5.

	Cutoff points (category values)			
	5	4	3	2
<i>FPR</i>	0.0345	0.2241	0.3296	0.4310
<i>TPR</i>	0.6471	0.8627	0.9020	0.9412

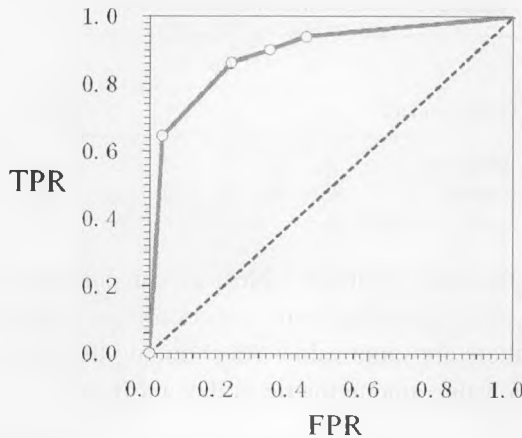


Fig. 27.2 The empirical *ROC* curve for the data in Table 27.5.

27.3.2 Plotting *ROC* curve

Let *FPR* be *x*-axis, *TPR* be *y*-axis, an empirical *ROC* curve can be constructed by plotting all operating points and joining the adjacent points by a straight line. Any *ROC* curve certainly includes the points (0, 0) and (1, 1). The two points correspond to the operating point of $Sen = 0, Spe = 1$ and of $Sen = 1, Spe = 0$ respectively.

For the data in Table 27.5, the empirical *ROC* curve is shown in Fig. 27.2. In theory, a worthless diagnosis, $TPR = FPR$, is a diagonal line from the origin to the top right corner which is usually called chance line. *ROC* curve usually lies over the chance line. The farther *ROC* curve is from the chance line, the better is the accuracy of a diagnostic test. An almost perfect test shows that its *ROC* curve goes straight up from the origin to nearly the top left corner, and then moves to the top right corner.

27.3.3 Calculating the area under ROC curves

The area under ROC curve denoted with AUC or A_Z , can reflect the accuracy of a diagnostic test. Its value ranges from 0.5 to 1. For a worthless test, $A_Z = 0.5$, for a perfect test, $A_Z = 1$. In general, $A_Z = 0.50$ – 0.70 , indicates a poor accuracy; 0.70 – 0.90 , a good accuracy; above 0.90 , an excellent accuracy (Swets, 1988). There are mainly three algorithms for A_Z and its standard error, one parametric method of bi-normal model and two nonparametric methods. Among them, the nonparametric method (Hanley and McNeil, 1982; 1983) is simpler and easier to understand. By a numerical example the method is introduced as the follows:

Suppose that there are n_a subjects with the values x_{a_i} ($i = 1, 2, \dots, n_a$) in the abnormal group and n_n subjects with the values x_{n_j} ($j = 1, 2, \dots, n_n$) in the normal group. It can be proved that the area under ROC curve (A_Z) is the probability that the test result in the abnormal group is larger than that in the normal group. The expressions are

$$A_Z = \frac{1}{n_a n_n} \sum_{j=1}^{n_n} \sum_{i=1}^{n_a} \psi(x_{a_i}, x_{n_j}), \quad (27.10)$$

$$\psi(x_{a_i}, x_{n_j}) = \begin{cases} 1 & x_{a_i} > x_{n_j}, \\ 0.5 & x_{a_i} = x_{n_j}, \\ 0 & x_{a_i} < x_{n_j}. \end{cases} \quad (27.11)$$

Equation (27.11) means that comparing each value x_{a_i} in the abnormal group with each value x_{n_j} in the normal group, if the former is larger than the latter the score is 1; if equal, the score is 0.5; otherwise the score is 0. Equation (27.10) gives the average score which is obtained from $n_a \times n_n$ comparisons, i.e., A_Z (Noted that if smaller test result indicates more abnormal, we should exchange the signs and $<$ in (27.11)).

The standard error of A_Z can be estimated by

$$SE(A_Z) = \sqrt{\frac{A_Z(1 - A_Z) + (n_a - 1)(Q_1 - A_Z^2) + (n_n - 1)(Q_2 - A_Z^2)}{n_a n_n}}, \quad (27.12)$$

where Q_1 is the probability that two randomly chosen subjects in abnormal group have larger test result than one randomly chosen subjects in normal group and Q_2 is the probability that one randomly chosen subject in abnormal group have larger test result than two randomly chosen subjects in normal group (assume that larger test result indicates more abnormal). The estimating method for Q_1 and Q_2 are displayed in Example 27.4.

By a hypothesis test

$$H_0 : A_Z = 0.5, \quad H_1 : A_Z > 0.5.$$

We can determine whether the difference between an obtained area under ROC curve and the area under chance line is statistically significant. The test statistics is

$$Z = \frac{A_Z - 0.5}{SE(A_Z)}.$$

When H_0 is true, it will approximately follow a standard normal distribution. The asymptotic $(1 - \alpha)$ confidence interval for A_Z can be constructed according to

$$A_Z \pm Z_\alpha SE(A_Z).$$

Example 27.4 41 persons with certain disease (abnormal group) and 193 persons with disease-free (normal group) were diagnosed by expert consultation. A research panel had used film data as a diagnostic test to all these subjects. They classified the film data into grade 1, 2, 3, 4, 5 according to their confidence degree to abnormal. The results are displayed in lines 1 and 2 of Table 27.7. How much is the accuracy of such a diagnostic approach?

Table 27.7 Calculating A_Z and its standard error related to the data of categorical variable.

Content	Categories					Total
	1	2	3	4	5	
1. Abnormal (x_a)	2	3	8	16	12	41 = n_a
2. Normal (x_n)	35	68	49	29	12	193 = n_n
3. $x_a >$ Category (y_a)	39	36	28	12	0	
4. $x_n <$ Category (y_n)	0	35	103	152	181	
5. $x_n y_a + x_n x_a / 2$	1400	2550	1568	580	72	6170
6. $x_n (y_a^2 + y_a x_a + x_a^2 / 3)$	56011	95676	50437	12219	576	214919
7. $x_a (y_n^2 + y_n x_n + x_n^2 / 3)$	817	15439	131651	444677	419772	1012356

Solution From Table 27.7 we can obtain the related parts for calculating the area under *ROC* curve A_Z and its standard error. Row 3 of Table 27.7 is to calculate the number of abnormal film above a category (i.e., the difference between the number of abnormal film (41) and the sum of frequency for the category and below). Row 4 is to calculate the number of normal film below a category (i.e., the sum of frequency below the category). The Rows 5, 6, 7 are calculated with the formulas listed in the first column of rows 5, 6, 7 accordingly. For example, the value for the cell of row 6 and category 2 is equal to

$$68(36^2 + 36 \times 3 + 3^2/3) = 95676,$$

$$A_z = \frac{\text{Total of row 5}}{n_n n_a} = \frac{6170}{193 \times 41} = 0.7797,$$

$$Q_1 = \frac{\text{Total of row 6}}{n_n n_a^2} = \frac{214919}{193 \times 41^2} = 0.6624,$$

$$Q_2 = \frac{\text{Total of row 7}}{n_n^2 n_a} = \frac{1012356}{193^2 \times 41} = 0.6629.$$

Substitute the values of A_Z , Q_1 , Q_2 , n_n and n_a into (27.11), we have

$$\begin{aligned} SE(A_Z) &= \{[0.7797(1 - 0.7797) + (41 - 1)(0.6627 - 0.7797^2) \\ &\quad + (193 - 1)(0.6629 - 0.7797^2)]/(41 \times 193)\}^{1/2} \\ &= 0.0403. \end{aligned}$$

The test statistics $Z = (0.7797 - 0.5)/0.0403 = 6.9304$, $P = 0.0000$. The 95% confidence interval for A_Z is equal to $0.7797 \pm 1.96 \times 0.0403$ or (0.7006, 0.8588), which does not include 0.5. Those results suggest that the accuracy of the film diagnosis is significantly different from 0.5.

Example 27.5 Calculate the area under *ROC* curve and its standard error for hypothetical continuous data in Table 27.3.

Solution For continuous *ROC* data, rank the test results for all subjects in two groups, make the test results as cutoff points (only use one value if there is a tie). The above method of calculating A_Z and its standard error can be applied. Table 27.3 is sorted into rows 1 and 2 in Table 27.8. Rows

Table 27.8 Calculating A_Z and its standard error related to the data of continuous variable.

Content	Cutoff points									Total
	1.77	4.61	5.01	6.24	8.29	10.22	12.87	13.35	15.90	
1. Case (x_a)	0	0	1	0	0	1	1	1	1	$5 = n_a$
2. Control (x_n)	1	1	0	1	1	0	0	0	0	$4 = n_n$
3. $x_a >$ cutoff point (y_a)	5	5	4	4	4	3	2	1	0	
4. $x_n <$ cutoff point (y_n)	0	1	2	2	3	4	4	4	4	
5. $x_n y_a + x_n x_a / 2$	5	5	0	4	4	0	0	0	0	18
6. $x_n (y_a^2 + y_a x_a + x_a^2 / 3)$	25	25	0	16	16	0	0	0	0	82
7. $x_a (y_n^2 + y_n x_n + x_n^2 / 3)$	0	0	4	0	0	16	16	16	16	68

3–7 in Table 27.8 can be obtained imitating those in Table 27.7.

$$A_Z = \frac{\text{Total of row 5}}{n_n n_a} = \frac{18}{4 \times 5} = 0.9,$$
$$Q_1 = \frac{\text{Total of row 6}}{n_n n_a^2} = \frac{82}{4 \times 5^2} = 0.82,$$
$$Q_2 = \frac{\text{Total of row 7}}{n_n^2 n_a} = \frac{68}{4^2 \times 5} = 0.85.$$

Substitute the values of A_Z , Q_1 , Q_2 , n_n and n_a into (27.11), we have

$$SE(A_Z) = \sqrt{\frac{0.9(1 - 0.9) + (5 - 1)(0.82 - 0.9^2) + (4 - 1)(0.85 - 0.9^2)}{5 \times 4}}$$
$$= 0.1118.$$

The test statistic $Z = (0.9 - 0.5)/0.1118 = 3.5777$, $P = 0.0003$. The 95% confidence interval for A_Z is (0.6809, 1), which does not include 0.5.

Note that: (1) The upper limit value of 95% confidence interval is written as 1 although it is 1.1191 from the calculation, which exceeds the possible maximal value 1; (2) In practical application of ROC analysis, the sample sizes for each group should be larger than 20, while the sample sizes of the above example for normal and abnormal groups are just 4 and 5 respectively, which is only for the convenience of demonstrating.

27.3.4 Comparing the areas under ROC curve

Assume there are two independent samples and hence two ROC curves and two areas under them. The following will introduce the comparison between the two areas under ROC curves.

To test the hypothesis of H_0 : the areas under two ROC curves are equal, firstly, the two areas under ROC curves and their standard errors should be calculated using Eqs. (27.11) and (27.12), denoted with A_{Z_1} , A_{Z_2} and SE_1 and SE_2 respectively; then use the following test statistic for comparison

$$Z = \frac{A_{Z_1} - A_{Z_2}}{\sqrt{SE_1^2 + SE_2^2}}. \quad (27.13)$$

When H_0 is true, it follows the standard normal distribution.

Example 27.6 For two diagnostic tests, the values of A_{Z_1} , A_{Z_2} , SE_1 and SE_2 are 0.8828, 0.9302, 0.0326 and 0.0264, respectively. Assume they are independently sampled. Is the difference between the two areas under ROC curves statistically significant?

Solution From Eq. (27.13), we have

$$z = \frac{0.8828 - 0.9302}{\sqrt{0.0326^2 + 0.0264^2}} = -1.1299.$$

The corresponding P value is 0.2585. This result suggests that the difference between the two areas under ROC curves is not statistically significant.

As for comparing the areas for two correlated diagnostic tests, one needs to estimate the correlation between two compared areas. The computation is rather complicated, and is not in the scope of this book.

27.4 Decision Making on Diagnostic and Screening Test

To assess whether a reliable and valid diagnostic test is cost-effective when it is put into practice, decision analysis is required to examine the characteristics of the test population, cost of the test, correct diagnostic benefit, risk of miss diagnosis, etc. The result of diagnostic or screening test can provide useful information for decision analysis.

Table 27.9 Clinical screening for skull fractures for emergency patients with head trauma.

True status	Screening		Total
	Positive	Negative	
Skull fractures	99	23	122
No skull fractures	2423	3305	5728
Total	2522	3328	5850

Example 27.7 The Royal College of Radiologists of UK conducted a study of 5850 patients with head injury and receiving skull radiography (Royal College of Radiologists, Costs and benefits of skull radiography for head injury. *Lancet* (1981) 2, 791–795). A clinical screening was done for each patient with the presence or absence of symptoms and signs possibly indicating the underlying brain damage before skull radiography. If some symptoms and signs were present, the patient was described as “clinical positive”. Skull fractures were confirmed by skull radiography. Table 27.9 shows 2522 clinical positives, with 99 skull fractures including vault fractures, depressed fractures, basal fractures and frontal, ethmoidal, or sphenoidal fractures; 3328 clinical negatives, with 23 skull fractures. A decision needs to be made whether a simple clinical screening is worthwhile for the patients with head injury or simply take all head injury patients radiographed.

Figure 27.3 shows a proposal expressed with a tree where only the patients with positive result in clinical screening are recommended to receive skull radiography. And as a contrast, Fig. 27.4 shows another decision tree where all patients are recommended to receive skull radiography. Decide which proposal is better.

Solution

- (1) Assumptions To simplify the process, let us make some assumptions first:
 - (i) For head injury without intracranial hematoma, mortality was zero.
 - (ii) Regardless of skull fracture, it is possible that patient has intracranial hematoma.

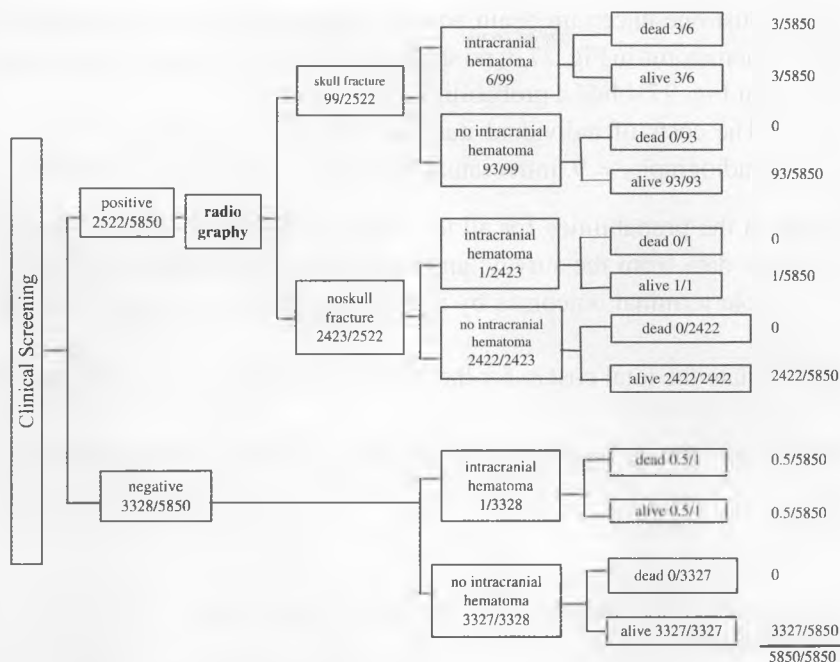


Fig. 27.3 The decision tree where only clinical screening positive patients receive skull radiography.

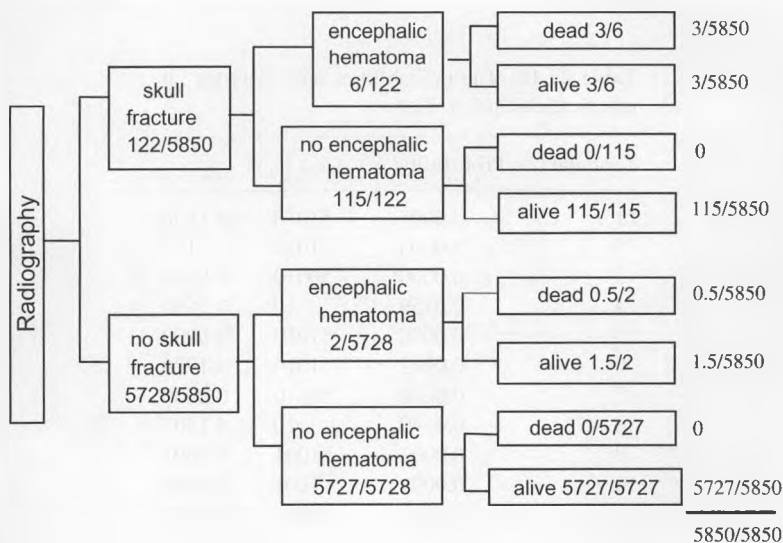


Fig. 27.4 The decision tree where all patients receive skull radiography.

- (iii) Just one uncertain death with screening negative and intracranial hematoma in Fig. 27.3; no skull fracture and intracranial hematoma in Fig. 27.4 had a probability of 0.5 for dead.
 - (iv) The costs of individual item are defined as: screening = 1, skull radiography = 9, intracranial hematoma = 1000, dead = 50000.
- (2) Fill in the probabilities for all the boxes in Figs. 27.3 and 27.4 based on the data from the survey; and calculate the probabilities p_i for all possible terminal outcomes by multiplying the probabilities along the paths.
 - (3) Calculate the total cost c_i for the i th terminal outcome by summing up the losing along the path, $i = 1, 2, \dots$
 - (4) Calculate the average losing (L) for Figs. 27.3 and 27.4 respectively,

$$L = \sum p_i c_i. \quad (27.14)$$

In Fig. 27.3, the probability of the first terminal outcome is $3/5850 = 0.0005$, the total cost is $c_1 = 1 + 9 + 1000 + 50000 = 51010$; The probability of the second terminal outcome is $3/5850 = 0.0005$, the total cost is $c_2 = 1 + 9 + 1000 = 1010$. For the probabilities of other terminal outcomes p_i , the total cost c_i and $p_i c_i$ are shown in Table 27.10.

Table 27.10 The probabilities and costs for all ramifications in Fig. 27.3.

Terminal (i)	Probability(p_i)	Cost (c_i)	$p_i c_i$
1	0.0005	51010	26.1590
2	0.0005	1010	0.5179
3	0.0000	50010	0.0000
4	0.0159	10	0.1590
5	0.0000	51010	0.0000
6	0.0002	1010	0.1726
7	0.0000	50010	0.0000
8	0.4140	10	4.1402
9	0.0001	51001	4.3591
10	0.0001	1001	0.0856
11	0.0000	50001	0.0000
12	0.5687	1	0.5687

Table 27.11 The probabilities and costs for all ramifications in Fig. 27.4.

Terminal (<i>i</i>)	1	2	3	4	5	6	7	8
Probability (p_i)	0.0005	0.0005	0.0000	0.0197	0.0001	0.0003	0.0000	0.9790
Cost (c_i)	51009	1009	50009	9	51009	1009	50009	9
$p_i c_i$	26.1585	0.5174	0.0000	0.1769	4.3597	0.2587	0.0000	8.8108

From Eq. (27.14), the average loss for Fig. 27.3 is

$$L_1 = 26.1590 + 0.5179 + \cdots + 0.5687 = 36.16. \quad (27.15)$$

Similarly, for each of the terminal outcomes in Fig. 27.4, the probability p_i , the total cost c_i and $p_i c_i$ are shown in Table 27.11.

From Eq. (27.14), the average loss for Fig. 27.4 is

$$L_2 = 26.1586 + 0.5174 + \cdots + 8.8108 = 40.28.$$

Comparing the average losses L_1 and L_2 , the proposal that only the patients with positive results in clinical screening receiving skull radiography is slightly better than the proposal that all patients receiving skull radiography without clinical screening. In fact, the cost of skull radiography considered above includes only the economic expense, the X-ray irradiation damage has not been taken into account. In addition to the low average lose, the proposal in Fig. 27.3 avoids X-ray irradiation for 3328 patients. From this viewpoint, the proposal in Fig. 27.3 is obviously better than the contrast proposal.

27.5 Computerized Experiments

Experiment 27.1 The relationship between PV_+ , PV_- and population prevalence Program 27.1 demonstrates the relationship between PV_+ , PV_- and population prevalence (P_0), when given sensitivity Sen and specificity Spe . In the program, $Sen = Spe = 0.95$; $P_0 = 0.1, 0.01, 0.001$ and 0.0001 , respectively; Lines 06 and 07 calculate the positive predictive value PV_+ and negative predictive value PV_- using Eqs. (27.6) and (27.7), respectively.

By running the program, we can get the results in Table 27.12.

Experiment 27.2 The area under ROC curve and its standard error for ordinal variable Program 27.2 is the SAS codes for Example 27.4.

Program 27.1 The relationship between PV_+ , PV_- and population prevalence.

Line	SAS languages and procedures	Line	SAS languages and procedures
01	DATA PREDICT;	07	PVNEG=SPE*(1-P0)/ (SPE*(1-P0)+(1-SEN)*P0);
02	S SEN=0.95;	08	OUTPUT;
03	S SPE=0.95;	09	END;
04	DO I=1 TO 4;	10	DROP I;
05	P0=10*(-I);	11	PROC PRINT;
06	PVPOS=SEN*P0/ (SEN*P0+(1-SPE)*(1-P0));	12	RUN;

Table 27.12 The results of Experiment 27.1.

OBS	SEN	SPE	P0	PVPOS	PVNEG
1	0.95	0.95	0.1	0.67857	0.99419
2	0.95	0.95	0.01	0.16102	0.99947
3	0.95	0.95	0.001	0.01866	0.99995
4	0.95	0.95	0.0001	0.00190	0.99999

For similar ordinal data, the program should be modified according to: (1) The data in lines 31 and 32: lines 31 and 32 are the frequencies of abnormal group and normal group, and sorted by categories from small to large respectively; (2) The number of categories K in line 02; (3) The program assumes that larger category indicates more positive test; if smaller category indicates more positive test, we should modify “ $YN(I)+LAG(XN(J))$,” to “ $YA(I)+LAG(XA(J))$,” in line 14; “ $YA(I)=NA-NA1(I)$,” to “ $YN(I)=NN-NN1(I)$,” in line 16; and “ $IF YN(I)=. THEN YN(I)=0$,” to “ $IF YA(I)=. THEN YA(I)=0$,” in Line 17. The program will output basic data, area under *ROC* curve, its standard error and 95% *CI*, *Z* test statistic, and other related middle results.

Experiment 27.3 The area under *ROC* curve and its standard error for continuous variable Program 27.3 displays partial SAS codes for Example 27.5. The aim of the program is to transform the layout of continuous data into ordinal data. To complete the calculation, lines 06 to 29 and line 34 in Program 27.2 should be appended to the end of this program.

Program 27.2 SAS codes for ROC analysis of ordinal variable.

Line	SAS languages and procedures
01	OPTIONS LINESIZE=70 PAGESIZE=MAX NODATE;
02	%LET K=5;/** K is the number of categories *****/
03	DATA F(KEEP=XN1-XN&K NN XA1-XA&K NA AREA Q1 Q2 SE_AREA Z P LCL95_A
04	UCL95_A YN1-YN&K YA1-YA&K AREAS1-AREAS&K Q1S1-Q1S&K Q2S1-Q2S&K);
05	INPUT XA1-XA&K XN1-XN&K @@;
06	ARRAY XN(*) XN1-XN&K; ARRAY XA(*) XA1-XA&K;
07	ARRAY NN1(&K); ARRAY NA1(&K);
08	ARRAY YN(&K); ARRAY YA(&K);
09	ARRAY AREAS(&K);
10	ARRAY Q1S(&K); ARRAY Q2S(&K);
11	NN=SUM(OF XN1-XN&K); NA=SUM(OF XA1-XA&K);
12	DO I=1 TO &K; DO J=1 TO I;
13	NN1(I)+XN(J); NA1(I)+XA(J);
14	YN(I)+LAG(XN(J)); /* YA(I)+LAG(XA(J)); *****/
15	END;
16	YA(I)=NA-NA1(I); /* YN(I)=NN-NN1(I); *****/
17	IF YN(I)=. THEN YN(I)=0; /*IF YA(I)=. THEN YA(I)=0; *****/
18	AREAS(I)=XN(I)*YA(I)+1/2*XN(I)*XA(I);
19	Q1S(I)=XN(I)*(YA(I)**2+XA(I)*YA(I)+1/3*XA(I)**2);
20	Q2S(I)=XA(I)*(YN(I)**2+XN(I)*YN(I)+1/3*XN(I)**2);
21	AREA=SUM(OF AREAS1-AREAS&K)/(NN*NA);
22	Q1=SUM(OF Q1S1-Q1S&K)/(NN*NA**2);
23	Q2=SUM(OF Q2S1-Q2S&K)/(NA*NN**2);
24	END;
25	SE_AREA=SQRT((AREA*(1-AREA)+(NA-1)*(Q1-AREA**2)+
26	(NN-1)*(Q2-AREA**2))/(NA*NN));
27	Z=(AREA-0.5)/SE_AREA; P=(1-PROBNORM(Z))*2;
28	LCL95_A=AREA-PROBIT(0.975)*SE_AREA;
29	UCL95_A=AREA+PROBIT(0.975)*SE_AREA;
30	CARDS;
31	2 3 8 16 12
32	35 68 49 29 12
33	;
34	PROC PRINT; RUN;

Modification of the program is similar to that of Experiment 27.2. Note that K in Program 27.3 denotes all possible cutoff points. To get the correct K value, we can first let K = the sample sizes of the diagnostic test, then run the SAS program and find the K value in the LOG window.

Program 27.3 Partial SAS codes for ROC analysis of continuous variable.

Line	SAS languages and procedures
01	OPTIONS LINESIZE=68 PAGESIZE=MAX NODATE;
02	DATA ROC; INPUT GP \$ NUMS;
03	DO I=1 TO NUMS; INPUT VALUE@@;
04	OUTPUT; END;
05	CARDS;
06	CASE 5
07	15.90 13.35 12.87 10.22 5.01
08	CONTROL 4
09	8.29 6.24 4.61 1.77
10	;
11	PROC FREQ ORDER=FORMATTED PAGE;
12	TABLES GP*VALUE /NOPRINT SPARSE OUT=A ;
13	DATA B;
14	SET A NOBS=KK; K=KK/2; PUT K=;
15	%LET K=9; /***** K is the number of cut-off point *****/
16	IF GP='CONTROL' THEN XNN=COUNT; ELSE XAA=COUNT;
17	PROC MEANS NWAY NOPRINT; VAR XNN XAA; CLASS VALUE;
18	OUTPUT OUT=C SUM=S1 S2 ;
19	PROC TRANSPOSE DATA=C OUT=D PREFIX=XN; VAR S1 ;
20	PROC TRANSPOSE DATA=C OUT=E PREFIX=XA; VAR S2 ;
21	DATA F; DROP _NAME_ NN11-NN1&K NA11-NA1&K I J;
22	SET D; SET E;

27.6 Practice and Experiments

1. To evaluate the accuracy of scrip method, which is a new method testing coli-group, zymotechnics is used as gold standard. The test results of a random sample from the population are given in Table 27.13. Calculate the sensitivity, specificity, Youden's index, positive (negative) likelihood ratio, and positive (negative) predictive value.
2. A hospital used marrow puncture as the gold standard for diagnosing iron deficiency anemia. The diagnostic results of blood assay for 100 patients in each of the case and control groups are displayed in Table 27.14. A researcher obtained the positive predictive value as 66/77 and the negative predictive value as 89/123 using Eqs. (27.8) and (27.9), respectively. Is his calculation method correct? Why? Calculate the predictive value if the population prevalence is 10%.

Table 27.13 The result for evaluation of the scrip method.

Zymotechnics	Scrip method		Total
	+	—	
+	91	8	99
—	5	36	41
Total	96	44	140

Table 27.14 The diagnostic results of blood assay for iron deficiency anemia.

Marrow puncture	Blood assay		Total
	+	—	
Case	66	34	100
Control	11	89	100
Total	77	123	200

3. Ventricular fibrillation (*VF*) is the primary reason for death of coronary heart disease. If ventricular premature beat (*VPB*) can be detected by cardiogram examination, *VF* can be prevented. *VPB* is tested by standard cardiogram (testing time is 1 minute). The “gold standard” was 24-hour cardiogram monitoring. 924 male patients with coronary heart disease received the test, and the results are displayed in Table 27.15 (Weiss NS, Clinical Epidemiology: The Study of the Outcome of Illness, Oxford Univ Press (1986), 28).

Let us define, for 24-hour cardiogram monitoring, *VPB* per hour < 10 as low risk for *VF*, and ≥ 10 as high risk; for standard cardiogram screening method, *VPB* = 0 as negative, and *VPB* ≥ 1 as positive. Perform the following calculations:

- (1) Calculate the sensitivity and specificity for standard cardiogram screening method and interpret their significance.
- (2) Calculate the positive predictive value and negative predictive value for standard cardiogram screening method and interpret their significance.

Table 27.15 Test results on *VPB* by standard cardiogram screening.

24-hours cardiogram (VPB per hour)	Frequency	Standard cardiogram	
		VPB≥1	VPB=0
0	444	0	444
1-9	247	12	235
10-49	120	40	80
≥ 50	113	79	34
Total	924	131	793

Table 27.16 The measures on mean corpuscular volume (*MCV*) of red blood cell for the patients and normal people.

Spinal puncture	Mean corpuscular volume (<i>MCV</i>) of red blood cell															
	52	58	62	65	67	68	69	71	72	72	73	73	74	75	76	77
abnormal	78	79	80	80	81	81	81	82	83	84	85	85	86	88	88	90
normal	60	66	68	69	71	71	73	74	74	74	76	77	77	77	77	78
	79	79	80	80	81	81	81	82	82	83	83	83	83	83	83	84
	84	84	84	85	85	86	86	86	87	88	88	88	89	89	89	90
	91	91	92	93	93	93	94	94	94	94	96	97	98	100	103	

See: J.R. Beck, E.K. Shultz, Arch Pathol Lab Med. 1986.

- (3) Calculate Youden’s index and likelihood ratio and interpret the significance of the positive likelihood ratio.
4. According to the above test results, assume that the mortality of *VF* is 10.0%, and let the cost be defined as: standard cardiogram=1, 24-hours cardiogram=30, death due to *VF* = 1000. Generate decision trees for the two *VF* prevention projects below and compare their average losses.

Project 1: 24-hour cardiogram monitoring for patients with positive result of standard cardiogram.

Project 2: 24-hour cardiogram monitoring for all patients with coronary heart disease.

5. Marrow diagnosis was used as the gold standard to confirm the diagnosis that 34 (out of 100) patients had iron deficiency anemia (abnormal group), while the other 66 patients were without iron deficiency anemia (normal group). The mean corpuscular volume (*MCV*) of red blood cell for patient was tested in advance, and shown in Table 27.16. Use *ROC* analysis to evaluate the accuracy of *MCV* in diagnosing iron deficiency anemia.

(1st edn. Yonyong Xu, Chuanhua Yu; 2nd edn. Yongyong Xu, Yi Wan)



Chapter 28

Design and Analysis of Sequential Experiments

28.1 Introduction

28.1.1 *Design of sequential experiment*

The experimental methods discussed in previous chapters have a common feature, that is, all of them have to determine the number of subjects (N) in advance, then assign subjects randomly to different groups and perform a statistical analysis after the data of all subjects are collected. This kind of experimental scheme pertains to those with fixed sample sizes. However, sequential trial has a different scheme, i.e., the number of subjects does not have to be pre-determined. In sequential trials, statistical analysis is performed when data of each subject are gathered. Once the conclusion of rejecting or not rejecting the null hypothesis is reached, the experiment could be stopped. This type of design is called sequential trial or experiment with unfixed sample size and the corresponding statistical analysis is called sequential analysis.

The sequential trials have three advantages. Firstly, it is more reasonable to treat sample size as a variable than as a constant because the sample sizes are decided by the total number of cases of certain diseases in the population and the entrance speed of subjects in clinical trials and epidemiological studies. Secondly, sequential analysis can make conclusions more quickly if the difference between treatment groups is substantial. Thus sequential trials potentially need fewer subjects and take shorter time than other types of experiments. Thirdly, since sequential trials perform statistical analysis at every time when the newly entered subjects have completed the trial, they can be immediately stopped once the difference between groups is found. In this sense, the sequential trials could be much more ethical than the trials with fixed sample sizes.

28.1.2 Group sequential design and interim analysis

The disadvantage of typical sequential trials is that they can only be applied to experiments in which the responses can be quickly measured and the entering time intervals are relatively short between two consecutive subjects. Thus, in recent years, researchers have paid more attention to group sequential design, which can be applied to the trials with mid-term or long-term follow-up times. The group sequential design provides the total number of stages (the number of interim stages plus a final stage) and a stopping criterion to reject, accept, or either reject or accept the null hypothesis at each interim stage. At each interim stage, all the data collected up to that time are analyzed, and the statistics and their associated standard error are computed. The test statistic is then compared with the critical values generated from the sequential design, and the trial is stopped or continued; in case a trial continues to the final stage, the null hypothesis is either rejected or accepted.

Interim analysis is also called “data-dependent stopping” or “early stopping”. Interim analyses are most often used to find convincing enough evidence to say that there is a statistically significant treatment difference, and that the difference is convincing enough to stop the trial at a time earlier than planned. In fact, the ethical and economic reasons are also taken into consideration to stop the trial early.

28.2 Design and Analysis of Sequential Trials

28.2.1 The qualitative responses

In this section we will discuss the Armitage scheme of sequential trial for quantal responses. Assuming that there are two treatments a_1 and a_2 , each subject receives the two treatments randomly. There are three possible results:

A: a_1 is superior to a_2 ,

B: a_2 is superior to a_1 ,

C: a_1 equals a_2 .

A and B are so-called preferences whose frequency of occurrence makes up n . Let π_1 be the response rate (e.g. efficacy rate) of a_1 and π_2 be the

Table 28.1 The boundaries of sequential trial with qualitative responses ($\theta_1 = 0.85$, $\alpha = \beta = 0.05$).

No. of test	No. of preferences (n)	Upper boundary (U)	Lower boundary (L)	False positive rate for individual test
1	7	7	-7	0.016
2	11	9	-9	0.022
3	14	10	-10	0.028
4	17	11	-11	0.033
5	20	12	-12	0.037
6	24	14	-14	0.038
7	26	14	-14	0.041
8	27	13	-13	0.047

response rate of a_2 , then we define a conditional probability as

$$\theta = \frac{\pi_1(1 - \pi_2)}{\pi_1(1 - \pi_2) + \pi_2(1 - \pi_1)}. \quad (28.1)$$

When $\pi_1 = \pi_2$, $\theta = 0.5$; when $\pi_1 > \pi_2$, $\theta > 0.5$; and when $\pi_1 < \pi_2$, $\theta < 0.5$. Thus, when we want to confirm whether or not a_1 is superior to a_2 , the hypothesis should take the form of $H_0 : \theta = 0.5$, $H_1 : \theta = \theta_1 (\neq 0.5)$. For a two-sided test $\alpha = \beta = 0.05$, we can plot the bound lines U , L , m' and m'' of the sequential trial diagram according to θ_1 and the number of unequal pairs n from Table 19 of Appendix II.

Example 28.1 Assuming a_1 is a certain kind of cough suppressants and a_2 is a placebo, we administer the two treatments to every patient in randomly. The effect of releasing cough is evaluated through the feelings of patients. Make an arrangement of the trials with sequential design when the accumulating data are analyzed at every new observation, and analyze the overall observations showed in Table 28.2.

Solution $H_0 : \theta = 0.5$; $H_1 : |\theta - 0.5| = 0.35$. Given two-sided significance level: $\alpha = 0.05$, $\beta = 0.05$, we can work out a sequential trial diagram (Fig. 28.1) using Table 28.1 from Table 19 of Appendix I.

The boundaries in Table 28.1 refer to critical values of eight repeated significance tests based on binomial distribution for overall level $\alpha = 0.05$. As the significance test is repeated, the false positive rate is going to increase, because the chance of mistaking a real effect for a large statistical fluctuation increases. For example, if one plans to have several repeated tests with a

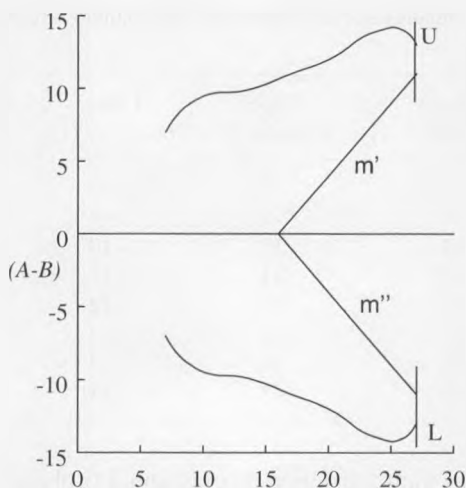


Fig. 28.1 The boundaries of sequential trial ($\theta = 0.85$, $\alpha = \beta = 0.05$).

planned alpha level of 5% for the observations from normal distribution, then the false positive rate would be changed to 8% for the second test, and 11% for the third test, 13% for the fourth test, 14% for the fifth test, 19% for the tenth test, and 37% for the 100th test. This example demonstrates that α is spending with the number of repeated tests. So that the nominal significance value is used to reduce the α 's spending for individual test, see the right of Table 28.1.

In Fig. 28.1, the horizontal axis represents the number of preferences and the vertical axis represents the difference between the numbers of preferences for a_1 and a_2 . The upper boundaries on line U , the lower boundaries on line L , the line m' and m'' are all worked out using the arithmetic function (n, U) , (n, L) and (n, y) according to Table 19 of Appendix I. We mark the result of each preference one by one in the diagram, i.e., starting a zigzag line at the origin and moving the unit to the right and upwards for each a_1 preference, or one unit to the right and downwards for each a_2 preference. On one hand, $H_1 : \theta = 0.85$, a_1 is superior to a_2 , should be accepted when the zigzag line reaches the upper boundary line. On the other hand, $H_1 : \theta = 0.35$, a_2 is superior to a_1 , should be accepted when the zigzag line reaches the lower boundary line. If the zigzag line reaches the line m' or m'' , then $H_0 : \theta = 0.5$ should be accepted.

Table 28.2 Data of sequential trial for comparison of cough suppressants (a_1) and placebo (a_2).

No.	n	Efficacy	$d = A - B$	$y = \sum d$	No.	n	Efficacy	$d = A - B$	$y = \sum d$
1		—	0	0	23	14	B	-1	9
2	1	B	-1	-1	24	15	A	1	10
3		—	0		25	16	A	1	11
4	2	A	1	1	26	17	A	1	12
5	3	A	1	2	27		—	0	
6		—	0		28		—	0	
7		—	0		29		—	0	
8		—	0		30	18	A	1	13
9	4	A	1	3	31	19	A	1	14
10	5	A	1	4	32		—	0	
11		—	0		33	20	B	-1	13
12		—	0		34	21	A	1	14
13		—	0		35	22	A	1	15
14	6	A	1	5	36		—	0	
15	7	B	-1	4	37	23	B	-1	14
16		—	0		38		—	0	
17	8	A	1	5	39	24	A	1	15
18	9	A	1	6	40	25	A	1	16
19	10	A	1	7	41	26	A	1	17
20	11	A	1	8	42	27	A	1	18
21	12	A	1	9	43		—	0	
22	13	A	1	10	44		—	0	
					45	28	B	-1	17

Note: "A" represents preference for a_1 , "B" represents preference for a_2 , "—" represents no difference between two treatments.

In this example (Table 28.2), the zigzag line reaches the upper boundary line at the 26th patient, i.e., the 17th preference. H_1 should be accepted and the conclusion should be made that the efficacy of cough suppressants is superior to that of the placebo. The sequential observations of the trial were plotted in Fig. 28.2.

28.2.2 Quantitative responses

In this section, we will introduce the Schneiderman-Armitage method which deals with quantitative responses in sequential trials. Quantitative response refers to the difference between the output measurements from

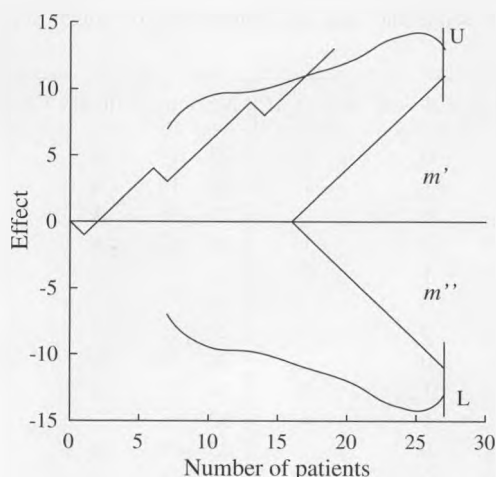


Fig. 28.2 Diagram of sequential trial for Table 28.2 ($\theta = 0.85$, $\alpha = \beta = 0.05$).

two treatments, i.e., $d_i \neq 0$ ($i = 1, 2, \dots, n$). In this kind of sequential trial each subject receives two treatments randomly, or each matched pair of patients receives one of the two treatments randomly. The hypothesis for this kind of design is

$$H_0 : \mu_d = 0, \quad H_1 : |\mu_d| = \delta, \quad \delta > 0.$$

The parameters δ and σ_d should be estimated from historical information, or be worked out as $\delta \approx |\bar{d}|/S_d$ and $\sigma_d \approx S_d$ using \bar{d} and S_d derived from pilot experiment. $\delta = |\mu_d|/\sigma_d$ is known as standardized effect size. Given $\alpha = \beta = 0.05$ and the value of δ , we can obtain the values of c_1 , b and (n', y') from Tables 20 and 21 of Appendix II, and then work out the upper and lower boundary lines and the right boundary line M for the diagram of sequential trial shown in Fig. 28.3. In Fig. 28.3, the horizontal axis represents the number of measurement pairs and the vertical axis represents the cumulative sum of differences between measurement pairs, i.e., $y = \sum d$. The upper and lower boundary lines are determined by Eqs. (28.2) and (28.3).

$$U : y = c_1 \sigma_d + b n \sigma_d, \quad (28.2)$$

$$L : y = -c_1 \sigma_d - b n \sigma_d. \quad (28.3)$$

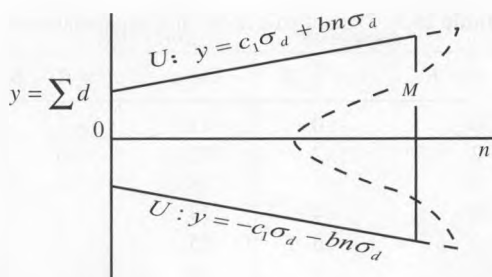


Fig. 28.3 Diagram for quantitative response sequential trial.

The right boundary line M is determined by Eq. (28.4).

$$M : (n = n' / \delta^2, y = y' \sigma_d / \delta). \quad (28.4)$$

In Eq. (28.4), (n', y') can be obtained from Table 21 of Appendix II.

To draw the diagram of sequential trial, we start a zigzag line from the origin according to the difference between the two measurements of each treatment pair, then connecting the points $(n, \sum d)$. If the zigzag line reaches the upper or lower boundary line, H_0 would be rejected and it can be concluded that a_1 is superior or inferior to a_2 . Otherwise, if the zigzag line reaches the right boundary line, H_0 would not be rejected and it should not be concluded that the difference between a_1 and a_2 is statistically significant.

Example 28.2 To compare the effects in releasing pain after operations, two patients were matched in the same operation on the same day by gender, nearest age, and allocated each patient with drug A and drug B randomly after operation. The response was pain score measured from 0–10. Plan a trial with sequential design and analyze the observations shown in Table 28.3.

Solution Set $H_0 : \mu_d = 0$, $H_1 : |\mu_d| = 2$, the overall two-sided significance level $\alpha = 0.05$, $\beta = 0.05$, $\hat{\sigma}_d = 1.2$, Refer to Table 20 and Table 21 in Appendix II, $c_1 = 3.03$, $b = 0.60$

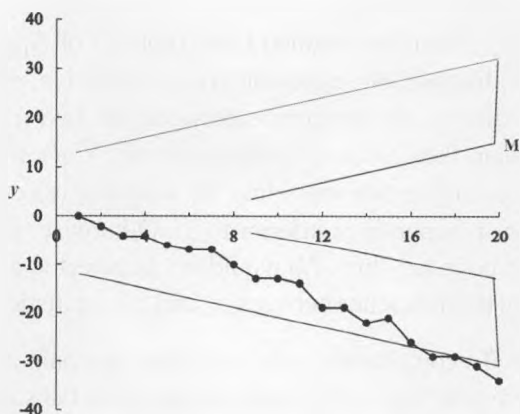
$$U : y = c_1 \sigma_d + bn \sigma_d = 3.03 \times 1.2 + 0.6 \times 1.2 \times n = 3.6 + 0.7 \times n,$$

$$L : y = -c_1 \sigma_d - bn \sigma_d$$

$$= -3.03 \times 1.2 - 0.6 \times 1.2 \times n = -3.6 - 0.7 \times n.$$

Table 28.3 Data of pain score in a sequential trial.

Days	$d = A - B$	$y = \sum d$	Days	$d = A - B$	$y = \sum d$
1	0	0	11	-1	-14
2	-2	-2	12	-5	-19
3	-2	-4	13	0	-19
4	0	-4	14	-3	-22
5	-2	-6	15	1	-21
6	-1	-7	16	-5	-26
7	0	-7	17	-3	-29
8	-3	-10	18	0	-29
9	-3	-13	19	-2	-31
10	0	-13	20	-3	-34

**Fig. 28.4** Diagram for quantitative response sequential trial for Table 28.3 ($\sigma_d = 1.2$, $\alpha = \beta = 0.05$).

Using Eq. (28.4) by setting $\delta' \approx 1.5/1.2 = 1$ and (n', y') in Table 21 of Appendix II, M in Fig. 28.4 was determined.

The differences of pain scores between the two measurements in each pair are listed in Table 28.3. Taking the operation days as the horizontal axis, $y = \sum d$ as the vertical axis represents the zigzag line was plotted in Fig. 28.4. In this example the zigzag line reached lower boundary at day 19, H_0 was rejected. The conclusion was that drug B would be better than drug A .

28.3 Group Sequential Schemes

In earlier sections of this chapter, we have discussed two typical sequential designs in which the subjects must be paired and enter the experiment sequentially. It could not be determined whether the experiment should be ended or continued until results of the present pair are obtained and analyzed. If it takes a long time to obtain the results for each subject, e.g. several weeks or several months, these kinds of sequential trials would become inappropriate. On other occasions, it is not practical to perform statistical analysis every time when results of each pair are obtained, so that what should be done is to perform statistical analysis when results are obtained during a certain period of time. For example, in a large multi-center clinical trial, experiments are performed simultaneously in several different regions but just one statistician is responsible for statistical analysis. The director of the trial wants to do statistical analysis, say every three months. In this case, group sequential trial methods can be useful. Group sequential trial, in which it is not essential for subjects to enter the experiment in pairs, can be employed where the collection of experiment results takes much longer time or statistical analysis must be repeatedly performed for several separate periods during the entire experimental process. Since group sequential trial has those advantages and can be applied with interim analysis, a term frequently appeared in the literatures, it could be widely used in clinical trials.

Group sequential trial was originally proposed by S. J. Pocock in 1977. The scheme is that the whole experiment is divided into a series of N stages. In each stage $2n$ subjects have been selected and are assigned to two treatment groups randomly, each of which has n subjects. Statistical analysis is performed for all the results from the first stage to the i th stage once the i th stage of experiment is completed, where $i = 1, 2, \dots, k$. If the null hypothesis is rejected in i th stage, the experiment can be terminated. Otherwise the experiment continues to the next stage. If the null hypothesis could not be rejected even in the last (k th) stage of the experiment, it should then be accepted.

28.3.1 Nominal significance level and test boundary

During the process of group sequential trial, repeated hypothesis testing is employed, which can increase the probability of type I error, i.e., the probability of incorrectly rejecting the null hypothesis could become larger.

In order to keep the type I error rate below the predetermined α , we must specify a lower α for the individual test at each stage. The significance level for each stage is so-called the nominal significance level and denoted by α' . Listed in Table 28.5 are a series of α' for different values of K when performing two-sided test for a normal distribution variable with known variance and significance level of 0.05. The critical value when α' is given is called test boundary. Also given in Table 28.5 are the test boundaries (Z') correspond to different α' 's from Pocock design.

The test boundaries frequently used in group sequential trial to test the efficacy of treatment A and treatment B are Pocock design, O'Brien-Fleming design and Peto design. Assuming they are normal distributions $N(\mu_A, \sigma^2)$ and $N(\mu_B, \sigma^2)$, and the overall significance level $\alpha = 0.05$, the test boundaries are determined as follows:

$$\text{Pocock: } B_i = \pm Z'_{\alpha_k} \quad (28.5)$$

$$\text{O'Brien-Fleming: } B_i = \pm Z_{\alpha/2} \sqrt{\frac{k}{i}}, \quad i = 1, 2, \dots, k \quad (28.6)$$

$$\text{Peto: } B_i = \begin{cases} \pm 3 & i < k, \\ \pm Z_{\alpha/2} & i = k, \end{cases} \quad (28.7)$$

where Z'_{α_k} is Pocock boundaries when given a nominal significance level in Table 28.5, $Z_{\alpha/2}$ is the standard normal deviate.

Among the above three methods for interim analysis, Pocock design requires the largest sample size to achieve specified power at the beginning of the study. O'Brien-Fleming design is very conservative, because the boundaries seem too large during the first stage. Peto design is in the middle between the other two, but more similar to O'Brien-Fleming design.

Example 28.3 The boundaries of interim analysis with two-sided test ($k = 4$, $\alpha = 0.05$) are given in Table 28.4.

28.3.2 Continuous responses

Supposing two treatments, A and B , are to be compared in a group sequential trial, the response variable is θ (e.g. θ could be the difference between two population means), which follows a normal distribution with known variance σ^2 . For a two-sided test, the null hypothesis is $H_0 : \mu = 0$, i.e.,

Table 28.4 The boundaries and nominal significance levels of interim analysis with two-sided test ($k = 4, \alpha = 0.05$).

No. of stages (k)	Pocock		O'Brien-Fleming		Peto	
	B_i	α'	B_i	α'	B_i	α'
1	2.36	0.0182	3.92	0.0000	3	0.0026
2	2.36	0.0182	2.77	0.0056	3	0.0026
3	2.36	0.0182	2.26	0.0238	3	0.0026
4	2.36	0.0182	1.96	0.500	1.96	0.0500

the efficacy of two treatments are equal, and the alternative hypothesis is $H_1 : \mu = \delta (\neq 0)$ or $H_1 : \mu = -\delta$, of which δ is the difference between the expected value of the response variable and 0. The significance level can be specified as $\alpha = 0.05$, and the test power $1 - \beta$. Given α and K , we can obtain α' using Table 28.5. Given the relationship

$$\Delta = \sqrt{n}\delta/\sigma$$

the sample size n for each stage can be worked out for different values of α and N .

Equation (28.8) shows a generalized equation about Δ . When the i th stage of experiment is completed, the estimate of θ , i.e., $\hat{\theta}$ and its variance $VAR(\hat{\theta})$ can be obtained, and we can then compute Δ as

$$\Delta = \sqrt{\frac{\delta}{i VAR(\hat{\theta})}}. \quad (28.8)$$

In group sequential trials of comparing two sample means with known variances, we further assume that the effects of two treatment groups A and B , i.e., x_A and x_B , follow normal distributions. Their conjunct variance σ^2 is known but the corresponding population means, μ_A and μ_B , are unknown. The response variable can be specified as the difference between the two means, i.e., $\theta = \mu_A - \mu_B$. Then the cumulative difference between the two means at i th stage is

$$\hat{\theta} = \bar{d}_i = \bar{x}_{Ai} - \bar{x}_{Bi}, \quad (28.9)$$

and

$$VAR(\hat{\theta}) = 2\sigma^2/(in).$$

Table 28.5 Parameters for design of group sequential trial ($\alpha = 0.05$).

Stages of experiment (k)	Nominal significance level (α')	Critical value (Z')	$\Delta = \sqrt{n}\delta/\sigma$	
			$1 - \beta = 0.90$	$1 - \beta = 0.95$
2	0.0294	2.178	2.404	2.664
3	0.0221	2.289	2.007	2.221
4	0.0182	2.361	1.763	1.949
5	0.0158	2.413	1.592	1.759
6	0.0142	2.453	1.464	1.617
7	0.0130	2.485	1.364	1.506
8	0.0120	2.512	1.282	1.415
9	0.0112	2.535	1.214	1.339
10	0.0106	2.555	1.156	1.275

According to the two equations above and Eq. (28.8), we can derive

$$\Delta = \frac{\sqrt{n}\delta}{\sqrt{2}\sigma} \quad (28.10)$$

so that we can compute n as

$$n = 2 \left(\frac{\sigma \Delta}{\delta} \right)^2. \quad (28.11)$$

When the experiment goes into i th stage ($i = 1, 2, \dots, k$), and denotes the cumulative difference between the two sample means as \bar{d}_i , statistical analysis can be performed using Z test.

$$Z_i = \frac{\sqrt{i}n\bar{d}_i}{\sqrt{2}\sigma}. \quad (28.12)$$

When $Z_i > Z'$ (Z' can be obtained from Table 28.5), the null hypothesis should be rejected, i.e., we can conclude that the effects of two treatments are different, and the experiment can be stopped. Otherwise the experiment should be continued. If the null hypothesis cannot be rejected even at the end of the N th stage, we cannot reject the null hypothesis for the whole experiment and should conclude that the effects of the two treatments are equal.

If the population variance is not known, the sample size of each stage of experiment, i.e., n , can be worked out through Eq. (28.11) using estimated variance in designing the experiment. On this occasion, t test should be

used to perform the statistical analysis and the degrees of freedom should be $2in - 2$. The null hypothesis should be rejected and the experiment should be stopped when we obtain the result of $P_i < \alpha'$.

The method mentioned above can be generalized to studies with binomial outcome variables.

Example 28.4 In a clinical trial comparing the efficacy of curing schizophrenia using a new drug A and a normal drug B , the difference in the total brief psychiatric rating scale (BPRS) score was used to evaluate the efficacy of the drugs. Given that the standard deviation of this variable for drug B was known to be 11, we expect that the score for drug A is higher than drug B by 4 units on average. To deal with this problem, we can perform a group sequential trial.

Solution What we should do first is to determine how many stages should we divide the whole experiment. For this example we specify $N = 5$, i.e., at most five times of repeated hypothesis tests will be performed, including 4 interim analyses and 1 analysis at the final stage. Second, we determine the confidence level of two-sided test for the whole experiment as $\alpha = 0.05$, and the power of test as $1 - \beta = 0.95$. Then we can obtain the necessary parameters using Table 28.5, i.e., $\alpha' = 0.0158$, $Z' = 2.413$, $\Delta = 1.759$. For this example, $\sigma = 11$, $\delta = 4$. According to Eq. (24.7), we can calculate the sample size for each stage, $n = 2(11 \times 1.759/4)^2 = 46.80$. As 47 subjects are needed in each group for each stage, the total sample size required is $47 \times 2 \times 5 = 470$.

Results at each stage for this example are listed in Table 28.6. The difference of sample means \bar{d}_i can be worked out by subtracting sample mean of group B for each stage \bar{X}_{Bi} from that of group A \bar{X}_{Ai} . Subsequently, Z_i can be worked out using Eq. (28.12).

$$Z_1 = \sqrt{1 \times 47} \times 3.3914 / (\sqrt{2} \times 11) = 1.495,$$

$$Z_2 = \sqrt{2 \times 47} \times 2.9014 / (\sqrt{2} \times 11) = 1.808,$$

$$Z_3 = \sqrt{3 \times 47} \times 3.8720 / (\sqrt{2} \times 11) = 2.956.$$

According to Table 28.5, $Z_1(1.495)$ and $Z_2(1.808)$ in the first two stages are all less than $Z'(2.413)$. Thus the null hypothesis H_0 should not be rejected and the experiment should be continued to the next stage. In the

Table 28.6 Results of group sequential trial comparing two sample means with known variance.

Stage (i)	Cumulative sample size	\bar{X}_{Ai}	\bar{X}_{Bi}	\bar{d}_i	Z_i	Results
1	94	37.3894	33.9980	3.3914	1.495	Not reject H_0 , experiment should be continued
2	188	36.1179	33.2165	2.9014	1.808	Not reject H_0 , experiment should be continued
3	282	36.3396	32.4676	3.8720	2.956	Reject H_0 , experiment should be stopped

third stage we obtain $Z_3(2.956) > Z'(2.413)$ such that the null hypothesis H_0 is rejected and the experiment should be stopped. At this time, we can conclude that the new drug A is preferable to drug B in terms of total BPRS score. The total sample size eventually used is 282.

The following equation can be used to calculate the sample size if we employ the parallel design method in this example.

$$n_A = n_B = 2 \left[\frac{(Z_{1-\alpha/2} + Z_{1-\beta})\sigma}{\delta} \right]^2$$

Given $\sigma = 11$, $\delta = 4$, $\alpha = 0.05$ and $1 - \beta = 0.95$, we obtain $n_A = n_B = 2[(1.960 + 1.645) \times 11/4]^2 \approx 196$, i.e., 196 subjects in each group and a total of 392 subjects are needed. Though the upper bound of the sample size for group sequential trial is bigger than the sample size for parallel design (e.g. 470 for the previous one and 392 for the latter one in this example), the actual sample size for group sequential trial is often smaller than that of parallel design if the two groups are different. Given $K = 5$, $\alpha = 0.05$ and $1 - \beta = 0.95$, the sample size for group sequential trial is $31.3(\sigma/\delta)^2$ on average. For this example, the actual sample size is $31.3 \times (11/4)^2 = 237$, with a reduction of 40% from that of the parallel design.

If the population variance is unknown in the previous example, we might estimate it at first in the stage of experiment designing. Assuming that $\sigma^2 = 112$, we can work out the same estimates of sample sizes as before.

Table 28.7 Results of group sequential trial comparing two sample means with unknown population variance.

Stage (i)	Cumulative sample size	$\bar{X}_{Ai} (S_{Ai})$	$\bar{X}_{Bi} (S_{Bi})$	t_i	P_i	Results
1	94	37.39 (11.65)	34.00 (11.19)	1.4394	0.1534	Not reject H_0 , experiment should be continued
2	188	36.12 (11.77)	33.22 (11.43)	1.7142	0.0882	Not reject H_0 , experiment should be continued
3	282	36.34 (10.92)	32.47 (11.24)	2.9346	0.0036	Reject H_0 , experiment should be stopped

The results for all stages are listed in Table 28.7, in which the t test for comparing two sample means is used to work out the P -values. As we can see in the table, the P -values for the first and second stages, i.e., $P_1(0.1534)$ and $P_2(0.0882)$, are greater than $\alpha'(0.0158)$, so that the null hypothesis could not be rejected and the experiment should be continued. In the third stage, $P_3(0.0074) < \alpha'(0.0158)$, so that the null hypothesis is rejected and the experiment should be stopped.

28.3.3 Discrete responses

Given the population probabilities for two treatment groups, A and B , be π_A and π_B , the response variable can be defined as the difference between two population probabilities, i.e., $\theta = \pi_A - \pi_B$. With $\bar{\pi} = (\pi_A + \pi_B)/2$, σ in Eq. (28.10) can be replaced by $[\bar{\pi}(1 - \bar{\pi})]^{1/2}$. Thus we can obtain

$$\Delta = \frac{\sqrt{n}(\pi_A - \pi_B)}{\sqrt{2\bar{\pi}(1 - \bar{\pi})}}. \quad (28.13)$$

Consequently, we can obtain

$$n = \frac{2\bar{\pi}(1 - \bar{\pi})\Delta^2}{(\pi_A - \pi_B)^2}. \quad (28.14)$$

When the experiment reaches the i th stage ($i = 1, 2, \dots, k$), a test based on approximate normal distribution can be used to test the difference between

the two sample rates.

$$Z_i = \frac{(P_{Ai} - P_{Bi})\sqrt{in}}{\sqrt{2\bar{P}_i(1 - \bar{P}_i)}} \quad (28.15)$$

For the equation above, P_{Ai} and P_{Bi} are cumulative sample frequencies for the two groups respectively, and $\bar{P}_i = (P_{Ai} + P_{Bi})/2$ is the pooled frequency. The null hypothesis can be rejected when $Z_i > Z'$, and the experiment can be stopped.

Example 28.5 The efficacy rates for a certain disease are to be compared between a new drug A and a normal drug B . The efficacy rate for drug B is already known as 0.7, we expect the efficacy rate for drug A could be raised to 0.85. Group sequential trial method can be used to solve the problem.

Solution What we should do first is again to determine the parameters, in this example $k = 5$, $\alpha = 0.05$ and $1 - \beta = 0.95$. From Table 28.5 we obtain $\alpha' = 0.0158$, $Z' = 2.413$ and $\Delta = 1.759$ and we can use Eq. (28.14) to estimate n when comparing two sample rates.

In this example, $\pi_A = 0.85$, $\pi_B = 0.7$, $\bar{\pi} = (0.85 + 0.7)/2 = 0.775$ and from Eq. (28.13) we obtain

$$n = 2 \times 0.775 \times (1 - 0.775) \times [1.759/(0.85 - 0.70)]^2 = 47.96.$$

Therefore, each of the two groups needs 48 subjects in each stage and the upper bound of the total sample size is $48 \times 2 \times 5 = 480$. For the i th stage ($i = 1, 2, \dots, k$), Eq. (28.15) can be used to test the difference between the two sample rates. The results of all stages of this example are listed in Table 28.8.

We see from Table 28.8 that at stage 2 $Z_2(3.346) > Z'(2.413)$, the experiment can be stopped and we can conclude that the efficacy rate of

Table 28.8 Results of group sequential trial comparing two sample rates.

Stage (i)	Cumulative sample size	P_{Ai}	P_{Bi}	Z_i	Results
1	96	0.8333	0.6875	1.674	Not reject H_0 , experiment to be continued
2	192	0.8854	0.6875	3.346	Reject H_0 , experiment to be stopped

drug A is higher than that of drug B . The actual sample size for this example is 192. If we use the parallel design method with $\pi_A = 0.85$, $\pi_B = 0.7$, $\alpha = 0.05$ and $1 - \beta = 0.95$, each group needs 199 subjects and the total sample size is 398, which is larger than the actual sample size of group sequential trial.

As for the determination of stage number k , factors such as time needed in clinical trials and how the subjects are selected should be taken into account. However, the stage number should not be larger than 10 because increasing k scarcely reduces the average sample size for stopping the experiment. In fact, the average sample size is not apparently reduced when increasing the stage number if it is larger than 5. Therefore, the stage number is usually set to be smaller than 5 unless the expected difference between the two treatment groups is much larger and the experiment is expected to end in early stages.

Although values of α' and Δ in Table I of Appendix II are derived from normal distribution with known variance, they can also be used in other cases such as normal distribution with unknown variance or binary variable etc. for which the error is very small.

28.4 Computerized Experiments

Experiment 28.1 Simulating experiment for group sequential trial To investigate the effect of extra vitamin intake of pregnant women on blood calcium density of newborn babies, we divide participating pregnant women into two groups, one given extra vitamin D and the other as control. In this study we divide the whole experiment into six stages. Based on clinical knowledge, the effect is regarded as clinically significant when the average rise of blood calcium density of the drug group from that of the control group is 0.3 mg% or higher. We also know that the standard deviation of blood calcium density of newborn baby is 1.2 mg%. Using the above parameters, Program 28.1 simulates a group sequential trial with a quantitative response and performs a Z -test on it.

Using a computer program, we generate two groups of equal number of blood calcium density measurements of newborn babies from two normal distributions respectively. The parameters used are $N(0.8, 1.2^2)$ and $N(1.1, 1.2^2)$, and a Z -test is also performed. The value k in line 02 of

Program 28.1 Simulation experiment for group sequential trial.

Line	Program	Line	Program
01	DATA GSEQU;	20	DDD=DDD+DD;
02	K=6;	21	D=DDD/I;
03	SD=1.2;	22	Z=(SQRT(I*N)*D)/(SQRT(2)*SD);
04	M1=0.8; M2=1.1;M=M2-M1;	23	OUTPUT;
05	ARRAY GSSPT(27) Z2-Z10	24	IF ABS(Z)> Z0 THEN STOP;
	AL2-AL10 DE2-DE10;		
06	INPUT Z2-Z10 AL2-AL10	25	END;
	DE2-DE10@@;		
07	Z0=GSSPT(K-1);	26	CARDS;
	ALPHA=GSSPT(K+8);		
08	DELTA=GSSPT(k+17);	27	2.18 2.29 2.36 2.41 2.45 2.48
09	N=CEIL(2*(SD*DELTA)**2/	28	2.51 2.54 2.56 0.0294 0.0221
	M**2);		
10	NN=2*N;	29	0.0182 1.0158 0.0142 0.0130
11	SUM1=0;SUM2=0;DDD=0;	30	0.0120 0.0112 0.0106 2.66 2.22
12	DO I=1 to K;	31	1.95 1.76 1.62 1.51 1.42 1.34 1.28
13	DO SUBJECT=1 TO N;	32	;
14	X1=M1+SD*NORMAL(0);	33	PROC PRINT NOOBS;
15	SUM1=SUM1+X1;	34	TITLE 'AN EXAMPLE
			OF GROUP
16	X2=M2+SD*NORMAL(0);		SEQUENTIAL DESIGN';
17	SUM2=SUM2+X2;	35	VAR K NN I SD D Z Z0 ALPHA;
18	END;	36	RUN;
19	DD=(SUM1-SUM2)/N;		

Program 28.1 can be changed to any value between 2 and 10 depending on how many stages the trial has. In the output (Table 28.9), “*k*” is the total number of stages in the experiment, “*NN*” is the sum of sample sizes of two groups, “*I*” is the order of the experiment, “*SD*” is the population standard deviation, “*D*” is the average difference between the two groups at the *i*th stage in Eq. (28.11), “*Z*” is the value of statistic of Z-test calculated by Eq. (28.12), “*Z0*” is the bound of the statistic for statistical significance and “*ALPHA*” is the nominal significance level α' .

28.5 Practice and Experiments

- 1. In Fig. 28.1, what are the basis in drawing the upper and lower bound line *U* and *L*, line *m'* and *m''*. Given the hypotheses $H_0:\theta=0.5$,

Table 28.9 An example of group sequential design.

<i>K</i>	<i>NN</i>	<i>I</i>	<i>SD</i>	<i>D</i>	<i>Z</i>	<i>Z0</i>	ALPHA
6	168	1	1.2	-0.15748	-0.85047	2.45	0.0142
6	168	2	1.2	-0.26084	-1.99222	2.45	0.0142
6	168	3	1.2	-0.46250	-4.32631	2.45	0.0142

$H_1: |\theta - 0.5| = 0.25$, draw a diagram of sequential trial, including the bound lines U , L , m' , m'' . Then use this diagram to perform a sequential trial for the data in Table 28.2 and observe the similarities and differences between the results of this and those of Fig. 28.2.

- According to a pilot experiment, the average difference between the observed values of two different treatments on the same subjects is $\bar{d} \approx 1.5$ and the standard deviation is $\sigma_d \approx 1.25$. Given significance level of two-sided test $\alpha = 0.05$ and $\beta = 0.05$, draw a diagram of sequential trial of quantitative response including the upper and lower bound line and the right bound line M . If $\bar{d} \approx 1.75$ and $\sigma_d = 1.25$, what will be the difference between the diagrams of sequential trial drawn in this case and the previous ones? Why?

Table 28.10 Results of sequential trial comparing the effects of two combinations of barium meals in the barium meal exam.

No. 2 of Shanghai with BCS		No. 2 of Shanghai with BL	
Patient No.	Effects evaluation	Patient No.	Effects evaluation
1	Preferable	1	Preferable
2	Preferable	2	Preferable
3	Preferable	3	Preferable
4	Preferable	4	Preferable
5	Preferable	5	Inferior
6	Preferable	6	Preferable
		7	Preferable
		8	Preferable
		9	Inferior
		10	Inferior

Source: Yang Shuqin, Guo Zuchao, Chinese medical encyclopedia (medical statistics).

Table 28.11 Measurements of blood cholinesterase activities of patients of a certain chronic disease and normal people (international units).

No. of matched pairs (n)	Cholinesterase activities of patients (x_1)	Cholinesterase activities of normal people (x_2)	Differences ($d = x_1 - x_2$)	$y = \sum d$
1	43.28	42.36	0.92	0.92
2	52.60	52.40	0.20	1.12
3	33.32	32.40	0.92	2.04
4	42.72	42.52	0.20	2.24
5	52.38	53.04	-0.66	1.58
6	53.64	52.64	1.00	2.58
7	52.98	52.56	0.42	3.00
8	34.40	32.40	2.00	5.00
9	42.54	42.29	0.25	5.25
10	43.00	42.51	0.49	5.74

3. To compare the effect of barium meal exam using two different combinations of barium meals, i.e., No. 2 of Shanghai with BCS (Barescoat-s) and No. 2 of Shanghai with BL (Barytgende Luxe), we perform a matched comparison after administering the two interventions to the same subjects. Given $\theta_1 = 0.95$ and two-sided significance level of $\alpha = \beta = 0.05$, compare the effects of two combinations of barium meals in the barium meal exam by drawing a diagram of sequential trial for the data listed in Table 28.10. If the effects are different, which is preferable?
4. The following table contains measurements of blood cholinesterase activities of patients of a certain chronic disease and normal people using slip semi-quantitative method. Draw a diagram of sequential trial according to the results of the study given in Table 28.11 and briefly explain the results.
5. Design a group sequential trial with 9 stages ($k=9$) to solve the problem in Experiment 28.1.

(1st edn. Yongyong Xu, Fubo Xue; 2nd edn. Yongyong Xu, Yi Wan)

Chapter 29

Systematic Review of Medical Research and Meta-Analysis

The systematic review on medical research is a basic and important step in medical studies. A good piece of review which summarizes the results of medical researches on a certain topic during a period, can evaluate the significance of the results, find some problems in those studies, and point out the further direction of investigation. With the development of evidence-based medicine, the synthesized investigation of medical researches becomes a bridge, by which the outcomes of medical researches in literatures can be well transformed into clinical practice by doctors. Traditional review of medical researches mainly depended on the authorities, who summarized and reviewed over the studies according to their own cognition in certain realm and their own understanding of the related courses. Different data collected by researchers with different experiences and different subjective perceptions might sometimes lead to totally different conclusions on the same topic. Obviously, traditional synthesized investigation of medical literatures lacks objectivity, and can hardly synthesize the outcomes of major researches quantitatively. In 1976, G.V. Glass developed methods for systematic review of research based on the mergence of reported statistics, and named as meta-analysis. Now meta-analysis has become a powerful tool for systematic review, especially for evidence-based medicine.

29.1 Basic Notions

29.1.1 *The definition of meta-analysis*

Example 29.1 In clinical practice of hospital management, there was a debate on whether it was necessary to use psychological treatment to reduce

Table 29.1 Five clinical studies for the number of days stay in hospital.

Study no.	Treatment			Control			Combined s_i	Effect d_i	t	P -value
	n_{1i}	\bar{x}_{1i}	s_{1i}	n_{2i}	\bar{x}_{2i}	s_{2i}				
1	13	5.00	4.70	13	6.50	3.80	4.27	0.351	0.895	0.380
2	30	4.90	1.71	50	6.10	2.30	2.10	0.571	2.474	0.016
3	35	22.50	3.44	35	24.90	10.65	7.91	0.303	1.269	0.209
4	20	12.50	1.47	20	12.30	1.66	1.57	-0.127	-0.403	0.689
5	8	6.50	0.76	8	7.38	1.41	1.13	0.779	1.554	0.143

the number of days stay in hospital. Research results from the literatures were not consistent. Part of the results showed that using psychological treatment could reduce the number of days stay of patients in hospital; the other results showed that psychological treatment had no statistical significance (Table 29.1). The problems that the systematic reviews of medical researches need to answer were: (1) whether the psychological treatment was able to reduce the number of days stay of patients in hospital? How great the effect was? (2) When would it work well in reducing the number of days stay in hospital?

Example 29.2 In clinical practice of pediatrics, there was a debate on whether it was necessary to use glucocorticoid to prevent neonate from ARDS. Research results from the literatures were not consistent. Part of the results showed that using glucocorticoid could prevent neonate from ARDS and lower the neonatal mortality rate; the other results showed that the decline of mortality rate had no statistical significance (Table 29.2). The problems that the systematic reviews of medical researches need to answer were: (1) whether the glucocorticoid was able to raise precautions against neonate ARDS and lower the mortality rate? How great was the effect? (2) When would it work well in raising precautions against ARDS to neonate?

Traditional review of medical literatures would only draw a conclusion about whether “there is statistical significance”. In Example 29.1, traditional review of medical literatures might report that only 1 out of 5 studies demonstrated statistical significance, which could not show the validity of psychological treatment; and it might also report that 4 out of 5 showed

Table 29.2 14 clinical trials for preventing neonate from ARDS.

Research No.	a_i	b_i	c_i	d_i	OR	χ^2	P-value
1	36	496	60	478	0.5782	6.299	0.012
2	1	68	5	56	0.1647	3.348	0.067
3	14	117	20	117	0.7000	0.925	0.336
4	3	64	7	52	0.3482	2.343	0.126
5	8	48	10	61	1.0167	0.001	0.974
6	3	61	12	46	0.1885	7.225	0.007
7	1	70	7	68	0.1388	4.423	0.035
8	4	77	11	52	0.2456	5.955	0.015
9	32	339	34	338	0.9384	0.061	0.805
10	5	44	4	27	0.7670	0.139	0.710
11	7	114	13	111	0.5243	1.804	0.179
12	0	23	1	21	0.0913	1.069	0.301
13	9	31	11	31	0.8182	0.151	0.697
14	6	89	9	85	0.6367	0.687	0.407
Amount	129	1641	204	1543	0.5950	19.759	0.000

Note: a_i, b_i, c_i, d_i are elements in i th fourfold.

reductive effect on the number of days stay in hospital, only one showed opposite effect. Also in Example 29.2, only 4 out of 14 studies demonstrated statistical significance, which could not show the validity of glucocorticoid; and it might also report that 13 out of 14 showed protective effect, only one showed harmful effect. There existed serious contradiction on the effect of interest factor indeed. Obviously, traditional ways could not solve the above-mentioned debate, and provide a quantitative synthesized result. The synthesized researches should have different reliabilities in information separately, but traditional ways would synthesize medical literatures taking equal weight mechanically, regardless of experimental condition or sample size. Thus meta-analysis could be used to answer these questions and provide a quantitative synthesized result on the effect of interest factors.

The initial meaning of meta-analysis is to collect enough research results through the literatures, then sum up after statistical analyses. In 1985, L. V. Hedges defined meta-analysis as the rubric used to describe quantitative methods by combining evidence across studies. Another definition of meta-analysis given by Hugue in 1988 is: "the term 'meta-analysis' refers

to a statistical analysis which combines or integrates the results of several independent clinical trials, considered by the analyst to be 'combinable'". Obviously, the last definition is reasonable, because it not only clearly defines the analytic purpose of meta-analysis as to combine or integrate the existing results of independent researches, but also explains that meta-analysis specifically requires the data to be "combinable". This working definition clarifies an ambiguous thinking that one can use meta-analysis for any data.

29.1.2 *Effect magnitude*

In order to answer the question on the treatment effect in meta-analysis, we should define the effect magnitude at first. The great contribution of G. V. Glass' to meta-analysis is the induction of "effect size" (ES), which replaces the method of merging the P -values or of voting for whether "there is statistical significance", making a big step in the synthetical investigation of research literatures. Since "effect size" is easily thought as a common noun, some scholars propose to replace it with "effect magnitude", and take it as a proper term of statistics. Effect magnitude is a standardized statistic for reflecting the degree of relationship between treatment and its effect for each investigation.

< 0 negative effect

Effect magnitude = 0 no treatment effect

> 0 positive effect.

Several statistics of effect magnitudes have been commonly used such as the standardized difference between the experimental group and control group (the difference across two means divided by standard deviation of the control group), odds ratio and coefficient of correlation. As relative measures, the statistics of effect magnitude are not affected by different unit scales used in different studies such that the effect magnitudes from different researches can be contrasted and/or combined. In Example 29.1, the effect magnitude is the standardized difference across two mean hospital stay (the difference between control group and the psychological treatment group divided by the combined standard deviation across the two groups). In Example 29.2, the effect magnitude is the logarithm of odds ratio.

29.1.3 *Homogeneity test, fix effect model and random effect model*

As the first step of a meta-analysis, a homogeneity test among results from different studies should be carried out. If the homogeneity holds (such as the standardized differences across two mean hospital stay from five different studies trend consistently in Example 29.1, and the logarithms of odds ratio from 14 different studies trend consistently in Example 29.2), we would choose a fixed effect model to combine the effect magnitudes ignoring the demographic difference among studying populations (such as the patient characteristics from different studies in Examples 29.1 and 29.2). If the homogeneity does not hold, it is necessary to double check the research reports on their design, object and handling measures, and find out the factors that might influence the results, and make correction to the factors according to practical situation. For example, some studies having excessive drop-out patients should not be included in meta-analysis; or one can use stratified meta-analysis according to the characteristics of research populations etc. The random effect model is a popular choice to estimate the size of heterogeneity with corrected standard error to the final effect size, and followed by a weighted combination of effect size over different studies.

29.1.4 *The design and plan of a meta-analysis*

Just like any other medical researches, meta-analysis is not simply a technique of looking for several literatures and making statistical analysis to obtain the result. We must make a design strictly at the beginning of the study, establish research project in detail, and make sure the research results are reliable, repeatable and scientifically sound. A meta-analysis as a whole includes several aspects showed below:

29.1.4.1 *Definition of the problem*

The problem to be investigated should be clearly defined at first; the scope should not be too wide and too complicated. Several related details should also be considered such as whether there are disputes about some problems, and is there any hint to further research directions and practical significance. Back to Examples 29.1 and 29.2, the problem is whether

psychological treatment can reduce the number of days stay of patients in hospital and whether the glucocorticoid can prevent neonate from ARDS, of which the purpose is just for prevention, not diagnosis nor treatment.

After clearly defining the research problem, one needs to specify both inclusion and exclusion criteria (such as including only randomized control clinical trials, or also including non-randomized control clinical trials), the indices used for describing the results (such as the mean hospital stay or the difference across two mean hospital stay in Example 29.1, the death rate or death odds ratio in Example 29.2), effect magnitudes to be used (such as the standardized difference across two mean hospital stay in Example 29.1, the logarithm of odds ratio in Example 29.2), statistical methods and models (such as the method of combining odds ratio, random effect model), and how to report the results etc.

29.1.4.2 *Retrieval strategy*

In the design of meta-analysis, the retrieval tactics should be specified in advance, which consists of the keywords and their order, retrieval methods (by electronic literature database or by hand), literature types (only include treatise or still include personal correspondence without publication of data). In the interest of obtaining more comprehensive data, we would better search through all the available electronic literature databases. In Examples 29.1 and 29.2, the retrieval tool was an electronic database, and the scope of retrieving only covered the articles published in its collection.

29.1.4.3 *Assessing research literatures*

Meta-analysis is not bringing the collected literatures into analysis without choice. In order to obtain the appropriate weight, the validity and reliability of the literatures need to be evaluated one by one. There are many items concerned with validity, such as the representative of the sample, grouping of randomization, using of "blinding", setting of placebo, appropriateness of statistical methods applied etc. There are also many items concerned with reliability, such as the accuracy of measuring methods, sensitivity and specificity of indices for effectiveness, quality control in data collection etc.

29.1.4.4 *Data analysis*

The methods of statistical analysis need to be specified in the design stage, including the choice of effect magnitude, synthesis of effect, estimation of confidence interval, homogeneity test and the choice of models etc.

29.1.4.5 *Interpretation of the result and prospect of further research*

The purpose of meta-analysis is to get a certain conclusion by synthesizing enough results. In Example 29.2, the purpose is to answer the question from clinical doctors: Can we use the glucocorticoid to prevent neonate from ARDS? How much of the neonatal mortality rate might be lowered after using glucocorticoid. If the result of meta-analysis can not make a definite judgment yet, further investigation direction and prospect of that field should be advised.

29.2 Statistical Methods Commonly Used in Meta-Analysis

29.2.1 *Collecting and arranging the data*

First of all, meta-analysis is entirely based on the data collected through literature review. Due to the influence of individual reviews and understanding of the reviewer, the results collected could show great divergence from different reviewers. According to the requirement of meta-analysis, the measurement indices for continuous variables should at least include sample mean, standard deviation or variance, and sample size of every treatment group; the frequency-type indices for discrete variable should include odds ratio or relative risk, frequency, standard error of frequency or total number of individuals and cases or deaths in each group. The data of odds ratio or relative risk can also be represented by regression coefficients and their standard errors of logistic regression, Cox regression or Poisson regression.

In collecting such data, it is best to get the original data from authors and basic frequencies listed in literatures. As this is not always possible, the results reported in literatures should sometimes be rearranged to meet the needs of meta-analysis proposed.

For example, sometimes only means of two groups and the P -value of hypothesis test are reported in the literature and no standard deviations

or standard errors being required by effect size directly. Given the sample size of each group under comparison and the exact P -value, we can use the equations below to estimate the standard error and the corresponding variance.

$$S_{\bar{x}_1 - \bar{x}_2} = \frac{\bar{x}_1 - \bar{x}_2}{t_{p,v}}, \quad (29.1)$$

$$S_c^2 = \frac{S_{\bar{x}_1 - \bar{x}_2}^2}{\frac{1}{n_1} + \frac{1}{n_2}}. \quad (29.2)$$

In analogy, we can estimate the standard error of the frequency.

If the literature gives only the mean (or odds ratio) and confidence interval, but no standard error, and the size of the sample is big enough, we can use Eq. (29.3) to estimate the standard error,

$$S_{\bar{x}} = \frac{\mu_u - \mu_l}{2 \times 1.96}, \quad (29.3)$$

where μ_l and μ_u are the lower limit and upper limit of the confidence interval. In the same way we can get the standard error of odds ratio.

Some literatures collected might not report the results with the important confounding factors or biases being adjusted. Effect magnitude in these literatures needs to be adjusted in principle. However, this will not be introduced in detail here.

29.2.2 Data presentation in meta-analysis

Frequency tables and proper plots on the results of each research would give a complete impression to the whole profile. Through these intuitive approaches, readers may discover easily heterogeneity between the different research results. For example, research results in Examples 29.1 and 29.2 can be plotted as Figs. 29.1–29.4, respectively.

Figure 29.1 is a histogram of weighted frequency for d 's of five psychological treatment clinical studies. The weighted frequency equals to the weighted sum of frequency f_i , that is, $\sum w_i f_i$, where w_i is the same as that in Table 29.3. Figure 29.1 shows that most of d 's are 0.2–0.8 with a peak at 0.35. The frequency distribution has not shown obvious heterogeneity between different studies. Figure 29.2 shows that since most sample sizes are small, and the confidence intervals are wide, the overall negative result

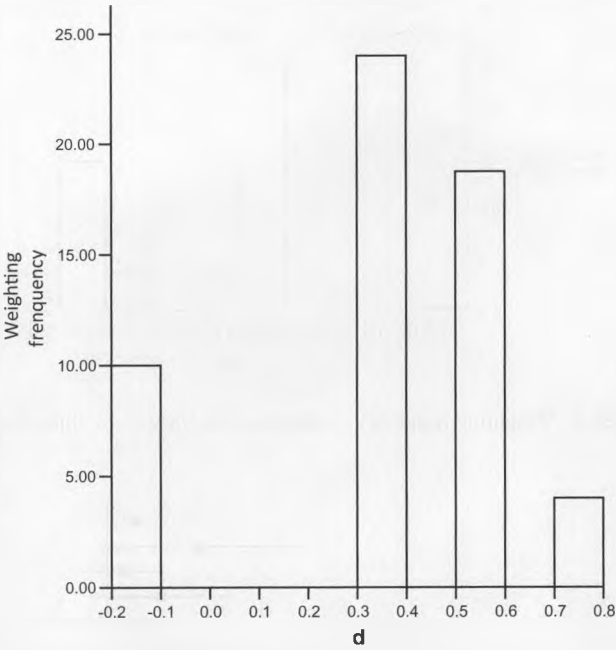


Fig. 29.1 Weighting frequency distribution for d 's of five psychological treatment clinical studies.

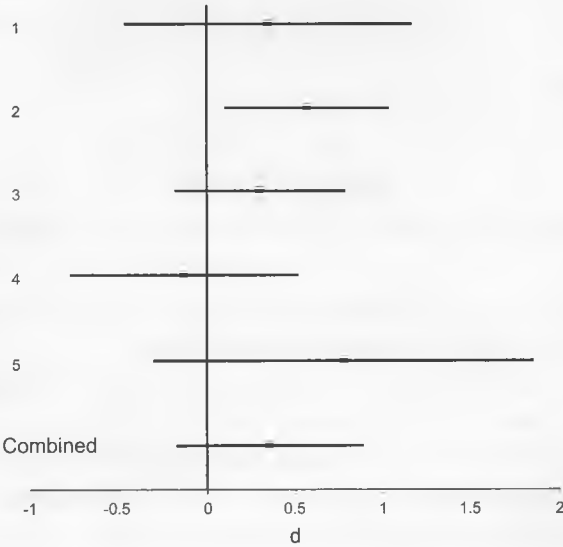


Fig. 29.2 95% confidence intervals for d 's of five psychological treatment clinical studies.

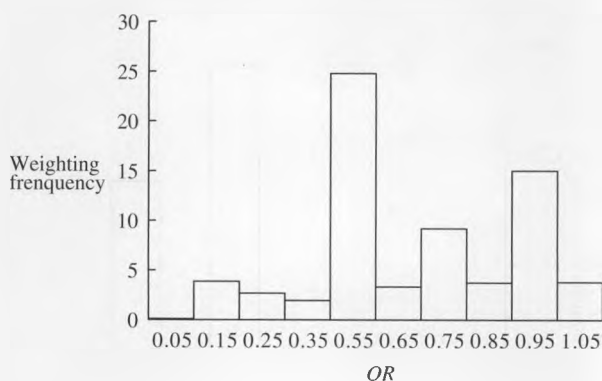


Fig. 29.3 Weighting frequency distribution for OR's of 14 clinical trails.

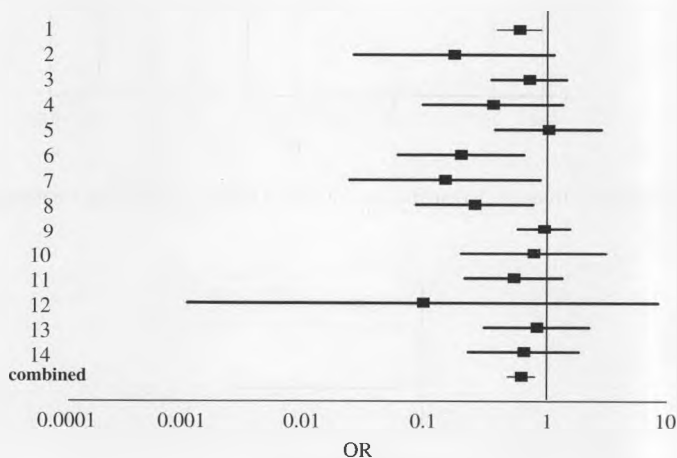


Fig. 29.4 95% confidence intervals for OR's of 14 clinical trails.

is obtained. Most 95% confidence intervals for d 's include 0, implying no effect of psychological treatment on hospital stay.

Figure 29.3 is a histogram of weighting frequency for ORs of 14 clinical trials. The weighted frequency equals the weighted sum of frequency f_i , that is, $\sum w_i f_i$, where w_i is the same as that in Table 29.5. Figure 29.3 shows that most of ORs are below 1, and a peak at 0.6. The frequency distribution has not shown heterogeneity between different studies. Figure 29.4 shows that since most sample sizes are small, and the confidence intervals are

wide, the overall negative result is obtained. Most OR's values are smaller than 1, implying a preventive effect of glucocorticoid.

29.2.3 Mergence of effect magnitudes

29.2.3.1 Weighted combination

In $k(\geq 2)$ researches, the means and variances of control group and experimental group in the i th study are denoted with \bar{x}_{1i} , \bar{x}_{2i} , s_{1i}^2 and s_{2i}^2 respectively; the combined variance of the two groups is denoted as s_i^2 ; and the effect magnitude is defined as

$$d_i = \frac{\bar{x}_{2i} - \bar{x}_{1i}}{s_i}, \quad i = 1, 2, \dots, k. \quad (29.4)$$

Suppose that the true effect for the i th study population is δ_i and the random effect is e_i , the observed effect size d_i can be expressed as the sum of the two in the following random effect model,

$$d_i = \delta_i + e_i, \quad i = 1, 2, \dots, k. \quad (29.5)$$

The estimate of the average effect size is

$$\bar{d} = \frac{\sum w_i d_i}{\sum w_i}, \quad (29.6)$$

where w_i is the weighting coefficient. Usually if there is no other information for the weight, the sum of sample size can be used as the weighting coefficient, $w_i = n_{2i} + n_{1i}$.

The observed variance of effect size among studies can be disentangled into two parts, the true variance of effect size and the variance of random error:

$$\text{Var}(d_i) = s_d^2 = s_\delta^2 + s_e^2$$

and s_d^2 can be estimated as

$$s_d^2 = \frac{\sum w_i (d_i - \bar{d})^2}{\sum w_i} = \frac{\sum w_i d_i^2 - \bar{d}^2 \sum w_i}{\sum w_i}. \quad (29.7)$$

Theoretically, the variance of random error can be calculated with the following formula

$$s_e^2 = \frac{4k}{\sum w_i} \left(1 + \frac{\bar{d}^2}{8} \right). \quad (29.8)$$

29.2.3.2 Homogeneity test

This is to compare the magnitude of the true variance of effect size with that of the variance of random error. The test result can lead to the choice of fixed effect model or random effect model. The null hypothesis is

$$H_0 : \delta_1 = \delta_2 = \dots = \delta_k.$$

We can calculate the following statistic

$$\chi^2 = \frac{k s_d^2}{s_e^2}. \quad (29.9)$$

Under H_0 , it follows a χ^2 distribution with $k - 1$ degrees of freedom. Based on the value of χ^2 , a P -value may be obtained. If $P \leq \alpha$, then H_0 is rejected, and a random effect model should be applied. Otherwise, a fixed effect model is applied.

29.2.3.3 The 95% confidence interval of mean effect magnitude

For a fixed effect model, the 95% confidence interval of mean effect magnitude is

$$\delta : \bar{d} \pm 1.96 s_{\bar{d}}, \quad (29.10)$$

where $s_{\bar{d}}$ is the standard error of \bar{d} ,

$$s_{\bar{d}} = \frac{s_e}{\sqrt{k}}. \quad (29.11)$$

For a random effect model, the 95% confidence interval of mean effect magnitude is

$$E(\delta) : \bar{d} \pm 1.96 s_{\delta}, \quad (29.12)$$

where

$$s_{\delta}^2 = s_d^2 - s_e^2. \quad (29.13)$$

Table 29.3 Five clinical studies for the number of days stay in hospital.

Study No.	Treatment			Control			Combined s_i	Effect d_i	Weighted combination		
	n_{1i}	\bar{x}_{1i}	s_{1i}	n_{2i}	\bar{x}_{2i}	s_{2i}			w_i	$w_i d_i$	$w_i d_i^2$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	13	5.00	4.70	13	6.50	3.80	4.27	0.351	6.50	2.282	0.801
2	30	4.90	1.71	50	6.10	2.30	2.10	0.571	18.75	10.706	6.113
3	35	22.50	3.44	35	24.90	10.65	7.91	0.303	17.50	5.303	1.607
4	20	12.50	1.47	20	12.30	1.66	1.57	-0.127	10.00	-1.270	0.161
5	8	6.50	0.76	8	7.38	1.41	1.13	0.779	4.00	3.116	2.427
Total									56.75	20.137	11.109

Example 29.3 In the synthetic investigation of medical literatures that were designed to study the association between psychological treatment and hospital stay, five clinical results were collected (Table 29.3). The reciprocal of combined variance for each study is used as the weighting coefficient. Work out a pooled weighted mean of effect magnitude based on the five studies.

Solution In Table 29.3, column (8) is the effect magnitude of various studies. Using Eq. (29.6) we calculate columns (9)–(11) to obtain $\sum w_i = 56.75$, $\sum w_i d_i = 20.137$, $\sum w_i d_i^2 = 11.109$. Using Eqs. (29.6) and (29.7), we have, $\bar{d} = 0.355$ and $s_d^2 = 0.0697$. With $k = 5$, we substitute them into Eq. (29.8) to obtain the error variance

$$s_e^2 = \frac{4 \times 5}{56.75} \left[1 + \frac{0.355^2}{8} \right] = 0.358.$$

Use Eq. (29.9) to carry out a homogeneity test, which results in $\chi^2 = (5)(0.0697)/0.358 = 0.973$, $\nu = 4$, $P > 0.50$. Therefore, we do not reject the hypothesis of $\delta_1 = \delta_2 = \dots = \delta_k$, and choose the fixed effect model for estimating the mean effect magnitude and the confidence interval.

By Eq. (29.11), standard error $s_{\bar{d}} = 0.2676$ is obtained; and by Eq. (29.10), the 95% confidence interval of mean effect magnitude is $0.355 \pm 1.96 \times 0.2676 = (-0.17, 0.88)$, which includes 0. Therefore, we do not reject the test hypothesis. It is concluded that the influence of psychological treatment on hospital stay could not be confirmed yet.

Table 29.4 Variation proportion of bone mineral density in ten studies.

Study No.	Drug A			Drug B			Combined s_i	Effect d_i	Weighted combination		
	n_{1i} (1)	\bar{x}_{1i} (2)	s_{1i} (3)	n_{2i} (4)	\bar{x}_{2i} (5)	s_{2i} (6)			w_i (9)	$w_i d_i$ (10)	$w_i d_i^2$ (11)
1	26	2.60	0.474	29	8.73	1.587	1.199	5.114	55	281.296	1438.682
2	27	2.41	0.639	32	3.94	1.541	1.216	1.259	59	74.258	93.462
3	28	1.40	0.639	27	6.20	1.574	1.193	4.023	55	221.281	890.276
4	21	3.58	0.144	23	5.39	1.209	0.881	2.055	44	90.435	185.874
5	26	2.22	0.277	22	7.54	1.246	0.866	6.141	48	294.774	1810.249
6	27	1.48	0.671	31	3.98	1.606	1.261	1.982	58	114.965	227.876
7	25	3.24	0.603	28	4.51	0.416	0.513	2.478	53	131.319	325.370
8	20	0.44	0.523	20	3.81	1.787	1.317	2.560	40	102.385	262.065
9	20	3.74	0.773	33	10.62	1.233	1.085	6.343	53	336.175	2132.333
10	28	1.89	0.942	30	6.70	1.132	1.045	4.604	58	267.039	1229.481
Total									523	1913.927	8595.668

Example 29.4 In the synthetic investigation of medical literatures that were designed to evaluate curative effect for osteoporosis between drug A and drug B, ten clinical RCT results of curative effect were collected (Table 29.4). The course of treatment was 12 months. The outcome variable is variation proportion of bone mineral density, which is a continuous variable. The standardized effect size is chosen as the effect magnitude. Sample size of each study is used as weighting coefficient. Work out a pooled weighted mean of effect magnitude based on the ten studies.

Solution In Table 29.4, column (8) is the effect magnitude of various studies. Using Eq. (29.6) we calculate columns (9)–(11) to obtain $\sum w_i = 523$, $\sum w_i d_i = 1913.926$, $\sum w_i d_i^2 = 8595.668$. Using Eqs. (29.6) and (29.7), we have, $\bar{d} = .6595$ and $s_d^2 = 3.0433$. With $k = 10$, we substitute them into Eq. (29.8) to obtain the error variance

$$s_e^2 = \frac{4 \times 10}{523} \left[1 + \frac{3.6595^2}{8} \right] = 0.2045.$$

Use Eq. (29.9) to carry out a homogeneity test, which results in $\chi^2 = (10)(3.0433)/0.2045 = 148.8056$, $\nu = 9$, $P < 0.01$. Therefore, we reject

the hypothesis of $\delta_1 = \delta_2 = \dots = \delta_k$, and choose to use the random effect model for estimating the mean effect magnitude and the confidence interval.

By Eq. (29.13), $s_\delta^2 = s_d^2 - s_e^2 = 3.0433 - 0.2045 = 2.8388$ is obtained, that is $s_\delta = 1.6849$; and by Eq. (29.12), the 95% confidence interval of mean effect magnitude is $3.6595 \pm 1.96 \times 1.6849 = (0.3572, 6.9618)$. It is concluded that the average proportion of improving bone mineral density by drug *B* is significantly higher than that by drug *A*, and is in the range of 0.36–6.96 times of standard error.

29.2.4 Mergence of odds ratio

29.2.4.1 Weighted combination

Suppose there are k studies, the i th result is

	Uncover factor		
	+	–	
Case	a_i	b_i	$OR_i = \frac{a_i d_i}{b_i c_i}$ $y_i = \ln(OR_i)$
Control	c_i	d_i	

Let μ_i and e_i be the population effect and random effect of y_i in the i th study respectively, and the random effect model is

$$y_i = \mu_i + e_i. \quad (29.14)$$

The weighted mean \bar{y}_w and its variance s_y^2 are

$$\bar{y}_w = \frac{\sum w_i y_i}{\sum w_i}, \quad (29.15)$$

$$s_y^2 = \frac{1}{\sum w_i}, \quad (29.16)$$

where w_i is the weighting coefficient,

$$w_i = \left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i} \right)^{-1}. \quad (29.17)$$

29.2.4.2 Homogeneity test

$$H_0 : \delta_1 = \delta_2 = \cdots = \delta_k,$$

$$Q = \sum w_i (y_i - \bar{y}_w)^2 = \sum w_i y_i^2 - \bar{y}_w^2 \sum w_i. \quad (29.18)$$

Under the null hypothesis, the statistic Q has a χ^2 distribution with $k - 1$ degrees of freedom. If $P \leq \alpha$, we choose the random effect model. Otherwise, we choose the fixed effect model.

29.2.4.3 The 95% confidence interval of OR

For the fixed effect model, the point estimate of OR and the 95% confidence interval of OR are

$$OR_c = \exp(\bar{y}_w). \quad (29.19)$$

For the random effect model, the weight w_i should be revised with w_i^* ,

$$w_i^* = \frac{1}{\frac{1}{w_i} + s_\mu^2}, \quad (29.20)$$

where s_μ^2 is the estimated variance of $\mu_i (i = 1, 2, \dots, k)$

$$s_\mu^2 = \frac{Q - k + 1}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}. \quad (29.21)$$

Using Eqs. (29.15), (29.16), (29.19) and (29.20), we re-calculate the weighted mean \bar{y}_{w^*} and variance s_y^2 . The point estimate of OR and the 95% confidence interval of OR are showed in Table 29.5.

Solution (of Example 29.2) Let us use the above procedure to analyze the data of Example 29.2 given in Table 29.2.

In Table 29.5, $\sum w_i = 68.4886$, and then we have $\sum w_i^2 = 780.45$. By substituting y_i and w_i from Table 29.5 into Eqs. (29.15), (29.16) and (29.18), we get $\bar{y}_w = -0.4833$, $s_y^2 = 0.0146$, $s_{\bar{y}} = 0.1208$, $Q = 13.9504$, $\nu = 13$, $P > 0.30$, so the fixed effect model may be chosen.

Based on the 14 studies, the point estimate of OR and the 95% confidence interval of OR are as follows:

$$OR_c = \exp(-0.4833) = 0.6167,$$

$$95\% \text{ CI of } OR = \exp(-0.4833 \pm 1.96 \times 0.1208) = 0.49-0.78.$$

Table 29.5 Meta-analysis of 14 clinical trials.

Research No.	a_i	b_i	c_i	d_i	OR_i	y_i	w_i
1	36	496	60	478	0.5782	-0.5478	20.5962
2	1	68	5	56	0.1647	-1.8036	0.8113
3	14	117	20	117	0.7000	-0.3567	7.2190
4	3	64	7	52	0.3482	-1.0549	1.9568
5	8	48	10	61	1.0167	0.0165	3.8135
6	3	61	12	46	0.1885	-1.6685	2.1988
7	1	70	7	68	0.1388	-1.9749	0.8534
8	4	77	11	52	0.2456	-1.4042	2.6801
9	32	339	34	338	0.9384	-0.0636	15.0217
10	5	44	4	27	0.7670	-0.2652	1.9617
11	7	114	13	111	0.5243	-0.6457	4.2094
12	0	23	1	21	0.0913	-2.3936	0.0902
13	9	31	11	31	0.8182	-0.2007	3.7518
14	6	89	9	85	0.6367	-0.4515	3.3247
Total	129	1641	204	1543	0.5950	-0.5191	68.4886

Example 29.5 Work out a meta-analysis of ten case-control studies based on the data given in Table 29.6 (Zhao Ning *et al.*, Modern Preventive Medicine (1993) 20(1)).

Solution Substituting y_i and w_i from Table 29.6 into Eqs. (29.15) and (29.18), we obtain $\bar{y}_w = 2.1741$, $Q = 19.3323$, $\nu = 9$, $P < 0.05$. Therefore, the random effect model is chosen in this case.

Using Eqs. (29.21) and (29.20), we have $s_\mu^2 = 0.3484$ and w_i^* listed in Table 29.6. Substituting y_i and w_i^* from Table 29.6 into (29.15) and (29.16), we obtain $\bar{y}_{w^*} = 2.3130$, $s_{\bar{y}}^2 = 0.0679$, $s_{\bar{y}} = 0.2606$.

Based on the ten studies, the combined point estimate of OR and the 95% confidence interval are

$$OR_c = \exp(2.3130) = 10.11,$$

95% CI of $OR = \exp(2.3130 \pm 1.96 \times 0.2606) = 6.06-16.84$.

The above analytical method not only carries out homogeneity test and weighted combination of OR s in case-control study, but also carries out weighted combination of relative risk RR in clinical RCT or cohort study, for example, a_i and b_i are the numbers of subjects with positive outcome and with negative outcome in the trial group, and c_i and d_i are those in

Table 29.6 Meta-analysis for the studies on association between liver cancer and HBV infection rate.

	a_i	b_i	c_i	d_i	OR_i	y_i	w_i	w_i^*
1	105	62	2	45	38.1048	3.6403	1.8252	1.1157
2	91	58	8	41	8.0409	2.0845	5.6300	1.9010
3	51	27	5	48	18.1333	2.8978	3.6039	1.5977
4	97	85	5	17	3.8800	1.3558	3.5600	1.5890
5	97	79	9	27	3.6835	1.3039	5.8440	1.9248
6	38	41	8	51	5.9085	1.7764	5.1200	1.8392
7	27	19	3	41	19.4211	2.9664	2.2352	1.2566
8	67	96	2	42	14.6563	2.6849	1.8210	1.1141
9	110	66	2	46	38.3333	3.6463	1.8316	1.1181
10	51	35	3	17	8.2571	2.1111	2.2710	1.2678
Total	734	568	47	375	158.4190	24.4674	33.7418	14.7241

control group respectively, while the positive rates (or negative rates) of the two groups are very small, RR can be estimated approximately by $a_i d_i / b_i c_i$, and then homogeneity test and weighted combination can be carried out.

29.3 Notes

29.3.1 Mergence of effect magnitude

The statistics of effect magnitude are not affected by different unit scales used in different studies such that the effect magnitudes from different researches can be contrasted and/or combined. The basic idea of meta analysis is that the results (for example, the difference between two means, the difference between two rates, correlation coefficient, OR and RR) from different studies can be combined, the average of effect magnitude will be calculated, and then the reliable conclusion would be drawn.

In practice, taking the logarithm of odds ratio for two rates as the effect magnitude is most familiar in medical literatures though, taking the difference between two rates to show clinical curative effect directly might sometimes be more meaningful. The preference depends on the analytical objectives.

29.3.2 Homogeneity test

If $P \leq \alpha$ in homogeneity test, first of all, the reason for disaccord from different studies should be explored, for example, including and excluding

criterion of the observation objects might be different. After excluding a few ineligible studies, whether the random effect model still should be used should be re-considered.

29.3.3 *Publish bias and funnel plot*

The most important problem in meta-analysis is publication bias, for example, medical journals tend to publish “positive” results with $P < 0.05$. It was reported that the publication rate of “positive” results in clinical trials was about 77%, and the publication rate of “negative” results was only 42%. Therefore, meta-analysis based on published literatures tends to obtain “positive” synthetic results. It is demanded that for high quality meta-analysis, one should collect as much as possible the relevant studies. The international medical journals have claimed that important researches must complete their public registration before starting. This measure might be helpful in retrieving some unpublished “negative” results.

The funnel plot is most frequently used to identify publication bias. It is a scatter plot, with the sample size (or reciprocal of standard error of effect magnitude) as y-axis, the effect magnitude (or logarithm of effect magnitude) as x-axis (Fig. 29.5), and each point refers to a study report. In practice, the independent studies with small sample size are in majority, they have more chance to present extreme values of effect magnitude than those with large sample size, and hence the corresponding dots distribute symmetrically at one bottom of one funnel plot; on the other hand, the independent studies with large sample size are in minority, their accuracies of effect magnitude are high, and hence the corresponding dots more concentrate at the top of the funnel plot. The basic hypothesis of funnel plot is that with the increase of sample size, the accuracy of estimating the effect magnitude increases, the range of variation gradually reduces, and converges to a dot finally. The scatter plot looks like an inverted symmetry funnel, so-called funnel plot. If the funnel plot presents asymmetric, then the publication bias may exist. Drawing of a funnel plot needs more independent studies (usually greater than five). Figure 29.6 is the funnel plot of Table 29.5.

29.3.4 *Fail-safe number*

Assume there were N_α literatures with “negative” results being missed, and incorporating these “negative” results with the current data, the “positive”

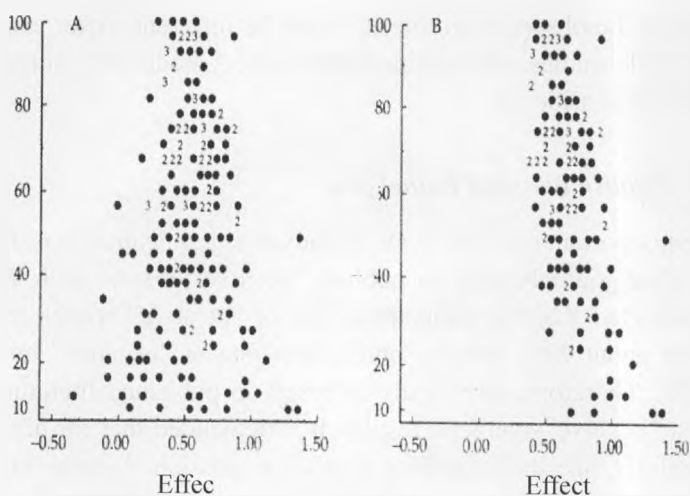


Fig. 29.5 A sketch map of funnel plot.

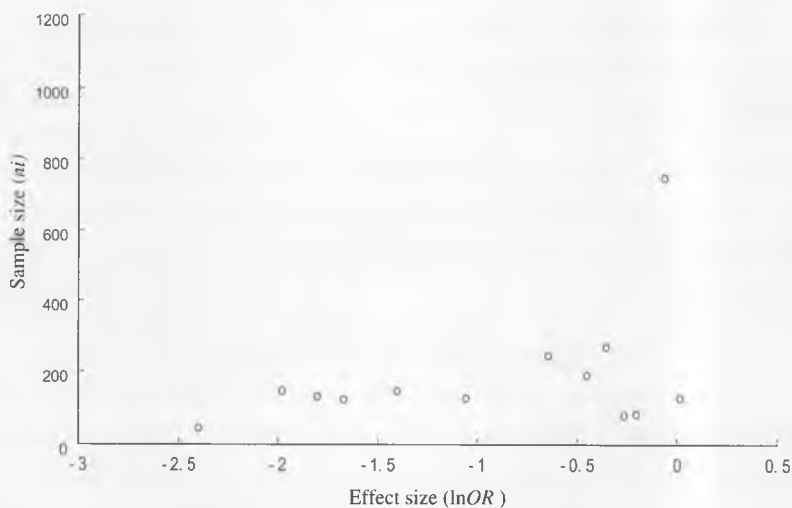


Fig. 29.6 Funnel plot of 14 clinical trails using glucocorticoid to prevent neonate from ARDS.

conclusion from current meta-analysis would be pulled down. This number is called as Fail-safe Number (N_f). It can be proved

$$N_f = \left(\frac{\sum Z_i}{Z_a} \right)^2 - k, \quad (29.22)$$

where Z_i is the i th standardized normal deviate obtained from the existing i th study for meta-analysis,

$$Z_i = \frac{\bar{X}_{1i} - \bar{X}_{2i}}{S_{\bar{X}_{1i} - \bar{X}_{2i}}}, \quad Z_i = \frac{\ln(OR_i)}{S_{\ln(OR_i)}} \quad \text{or} \quad Z_i = \frac{p_{1i} - p_{2i}}{S_{p_{1i} - p_{2i}}}$$

or transformed according to the P -value (one-sided probability) reported by the i th study through a table of standard normal distribution. N_α is the possible number of missed literatures in which effect magnitude is 0 at the test level α . The bigger N_α is, the smaller possibility is the current result being pulled down, and the smaller publication bias is.

As an example, for the merge of effect magnitude in Table 29.5, we have $\sum Z_i = -16.9834$, one-sided critical value $Z_{0.05} = 1.645$ is given, substituting into (29.22), we obtain

$$N_{0.05} = \left(\frac{\sum Z_i}{Z_{0.05}} \right)^2 - k = \left(\frac{-16.9834}{1.645} \right)^2 - 14 = 92.59 \approx 93.$$

The conclusion is that if the merging result $\hat{OR}_c = 0.6167$ of effect magnitude is only held due to missed literatures, in which effect magnitude is 0 ($OR = 1$) at α test level, the current result might be pulled down after adding 93 reports about failure to prevent neonate from ARDS by using glucocorticoid. If the reader thinks that the number 93 is big enough, and that is almost impossible to have so many "negative" literatures missed, then the publication bias might not be serious.

29.4 Computerized Experiments

Experiment 29.1 Merge of effect magnitudes Program 29.1 is for Example 29.4. Lines 01–19 are to input data, and calculate the combined standard deviation SSC and effect magnitude d_i . S_NUM is the number of studies; NTI XII STI NCI X2I and SCI represent the sample size, mean and standard deviation in drug A and drug B respectively. Lines 20–26 are to calculate w_i , $w_i d_i$, $w_i d_i^2$ in amounts as SWI, SWID and SWID2. Lines 27–39 are to calculate the weighted mean of the effect magnitudes and \bar{d} , and to perform the homogeneity test as well as to calculate 95% CI with the exact P -value of the homogeneity test, where AVD for \bar{d} , SD2 for s_d^2 , SE2 for s_e^2 , SDEL2 for s_δ^2 , SDBAR for $s_{\bar{d}}$, and CHISQ and P are the values of

Program 29.1 The meta-analysis of the differences between means.

Line	Program	Line	Program
01	DATA A;	21	ID= _N_;
02	INPUT NTI X1I STI NCI X2I SCI;	22	IF _N_ =S_NUM THEN DO;
03	SSC=SQRT(((NTI-1)*STI**2+(NCI-1)*SCI**2)/(NTI+NCI-2));	23	SWI=SWI1;SWID=SWID1;
04	DI=(X2I-X1I)/SSC;	24	SWID2=SWID21;
05	WI=NTI+NCI; WID=WI*DI;	25	END;
06	WID2=WI*DI**2;	26	PROC SORT;
07	S_NUM=10;	27	BY DESCENDING ID;
08	CARDS;	28	DATA C; SET B;
09	26 2.60 0.474 29 8.73 1.587	29	AVD=SWID/SWI;
10	27 2.41 0.639 32 3.94 1.541	30	SD2=(SWID2-AVD**2*SWI)/SWI;
11	28 1.40 0.639 27 6.20 1.574	31	SE2=4*S_NUM/SWI*(1+AVD**2/8);
12	21 3.58 0.144 23 5.39 1.209	32	CHISQ=S_NUM*SD2/SE2;
13	26 2.22 0.277 22 7.54 1.246	33	DF=S_NUM-1;
14	27 1.48 0.671 31 3.98 1.606	34	P=1-PROBCHI(CHISQ,DF);
15	25 3.24 0.603 28 4.51 0.416	35	IF SD2>SE2 THEN
16	20 0.44 0.523 20 3.81 1.787	36	SDEL2=SD2-SE2;
17	20 3.74 0.773 33 10.62 1.233	37	ELSE SDEL2=0;
18	28 1.89 0.942 30 6.70 1.132	38	LOW=AVD-1.96*SDEL2**0.5;
19	;	39	UP=AVD+1.96*SDEL2**0.5;
20	DATA B; SET A;	40	SDBAR=SE2**0.5/S_NUM**0.5;
	SWI1+WI;SWID1+WID;	41	FLOW=AVD-1.96*SDBAR;
	SWID21+WID2;	42	FUP=AVD+1.96*SDBAR;
		43	PROC PRINT;
		44	RUN;

χ^2 and P of homogeneity test. The terms AVD, FLOW, FUP, LOW and UP express the estimate of combined OR , the lower and upper limits of 95% CI of OR by the fixed effect model and the random effect model respectively. Lines 40–41 are to output the results.

Experiment 29.2 Mergence of OR s Program 29.2 is for Example 29.2. Line 01 is the selection of output format. Lines 02–23 are to input data, calculate OR and effect magnitude y_i ; S_NUM is the number of studies; NAI NBI NCI and NDI denote the cell numbers a , b , c and d respectively, of which a in the 11th study is supposed to be 0.1 rather than the initial value 0. Lines 26–37 are to calculate w_i , w_i^2 , $w_i y_i$, $w_i y_i^2$, \bar{y}_w , $s_{\bar{y}_w}^2$ (denoted

Program 29.2 The meta-analysis of OR data.

Line	Program	Line	Program
01	OPTIONS LS=74 PS=MAX NOCENTER NODATE;	29	ID= _N_;
02	DATA A;	30	IF _N_ =S_NUM THEN DO;
03	INPUT NAI NBI NCI NDI;	31	SW1=SWI1;SWI2=SWI21;
04	OR=(NAI*NDI)/(NBI*NCI);	33	SWIY1=SWIY11; SWIYI2=SWIYI21;
05	YI=LOG(OR);	33	YWBAR=SWIY1/SWI; SYBAR2=1/SWI;
06	WI=(1/NAI+1/NBI+1/ NCI+1/NDI)**(-1);	34	Q=SWIYI2-YWBAR**2*SWI;
07	NI=NAI+NBI+NCI+NDI;	35	SMU2=(Q-S_NUM+1)/ (SWI-SWI2/SWI);
08	S_NUM=14;	36	END;
09	CARDS;	37	P=1-PROBCHI(Q,S_NUM-1);
10	36 496 60 478	38	PROC SORT;
11	1 68 5 56	39	BY DESCENDING ID;
12	14 117 20 117	40	DATA C; SET B;
13	3 64 7 52	41	FORC=EXP(YWBAR);
14	8 48 10 61	42	FORLOW=EXP (YWBAR-1.96*SYBAR2**0.5);
15	3 61 12 46	43	FORUP=EXP (YWBAR+1.96*SYBAR2**0.5);
16	1 70 7 68	44	SMU2S+SMU2;
17	4 77 11 52	45	WIS=1/(WI**(-1)+SMU2S);
18	32 339 34 338	46	WISYI=WIS*YI;
19	5 44 4 27	47	SWIS1+WIS;SWISY11+WISYI;
20	7 114 13 111	48	IF _N_ =S_NUM THEN DO;
21	0.1 23 1 21	49	SWIS=SWIS1;SWISY1=SWISY11;
22	9 31 11 31	50	YBAR=SWISY1/SWIS;
23	6 89 9 85	51	SY2=1/SWIS;
24	;	52	ORC=EXP(YBAR);
25	DATA B; SET A;	53	ORLOW=EXP(YBAR-1.96* SY2**0.5);
26	WI2=WI**2;WYI=WI*YI; WYI2=WI*YI**2;	54	ORUP=EXP(YBAR+1.96* SY2**0.5);
27	SWI1+WI; SWI21+WI2;	55	END;
28	SWIY11+WYI; SWIYI21+WYI2;	56	PROC PRINT;
			RUN;

with SWI, SWI2, SWIY1, SWIY2, YBAR and SYBAR2), Q statistic of homogeneity test and the P -value of the test. Lines 38–54 are to calculate the estimate of combined OR and its 95% CI, where $WIS = w_i^*$, $YWBAR = \bar{y}_w$, $YBAR = \bar{y}$, $SMU2 = s_\mu^2$, $SY2 = s_y^2$; $FORC = O\hat{R}_c$, $FORLOW$, $FORUP$ and $ORC = O\hat{R}_c$, $ORLOW$, $ORUP$ express the estimate of combined OR , the lower limit and the upper limit of its 95% CI of the fixed effect model and of the random effect model respectively. Lines 55–56 are to output the results.

29.5 Practice and Experiments

- 1. Table 29.7 consists of the results of four studies that involve a psychological test about self-esteem. Try to perform a meta-analysis (Extracted from: Minghuang Hong, Chinese J. of Health Statistics (1992) 9, 1).
- 2. Work out a meta-analysis on the data of three case-control studies on HBV in Table 29.8.
- 3. Table 29.9 is the results of 4 RCTs in which peptic ulcer disease was treated by domestic Ranitidine in the trial group and by Cimetidine in

Table 29.7 Four studies on psychological test about self-esteem.

No.	Control		Training		Total SD
	n	X	n	X	
1	41	11	41	17	16
2	29	225	33	175	100
3	104	9	98	12	7
4	11	23	11	31	12

Table 29.8 Three case-control studies on HBV.

No.	HBV (+)		HBV (–)	
	Case	Control	Case	Control
1	44	17	12	39
2	25	12	21	80
3	55	10	14	128

Table 29.9 Curative effect of treating peptic ulcer disease.

Study no.	Healing rate (%)	
	Trial	Control
1	100.0 (7/7)	92.8 (13/14)
2	83.3 (30/36)	80.0 (20/25)
3	87.1 (54/62)	68.8 (44/64)
4	78.1 (25/32)	69.2 (18/26)

Table 29.10 Thickness of metacarpus II cortex of girl children in 11 investigations.

Study no.	High FI area			Fit FI area			Combined SD (7)
	N_{2i} (1)	X_{2i} (2)	s_{2i} (3)	n_{1i} (4)	x_{1i} (5)	s_{1i} (6)	
1	26	2.26	0.32	42	2.33	0.33	0.326
2	55	2.39	0.31	40	2.49	0.32	0.314
3	46	2.50	0.30	50	2.67	0.35	0.327
4	45	2.64	0.26	50	2.90	0.45	0.372
5	45	2.81	0.35	45	2.93	0.36	0.355
6	52	2.95	0.46	55	3.27	0.37	0.416
7	46	3.15	0.39	42	3.48	0.48	0.435
8	45	3.47	0.46	51	3.73	0.54	0.504
9	45	3.63	0.38	45	3.81	0.40	0.390
10	42	3.81	0.41	45	4.16	0.42	0.415
11	44	3.99	0.56	25	4.18	0.41	0.511

the control group. Work out a meta-analysis on the data of odds ratio for ulcer healing rates between domestic Ranitidine and Cimetidine.

4. Sample size of each study is used as weighting coefficient. Work out a meta-analysis on the data of 11 studies in Table 29.10.

(1st edn. and 2nd edn. Yongyong Xu, Changsheng Chen, Jiqian Fang)



Chapter 30

Comparative Effectiveness Research

30.1 Background

All countries in the world are facing the problem of limited resources for health care. Therefore it is important that clinical testing, treatment and prevention methods should be evaluated without bias, so as to provide medical professionals and health administrators information to select effective and low-cost methods of prevention or treatment and policies in health care. Although there are a lot of data reporting the effects of new medications, equipments, and medical procedures, strict comparison research between different treatments is rare. For example, two new medications are both better than the placebo, but which one is better? Such studies are lacking; and findings of such studies may bring pharmaceutical factories great risks because of the potential unfavorable results. New medications and equipments are emerging quickly and continuously though, whether the new ones are better than the old ones should be studied carefully.

At the same time, among existing data, there is a lack of evidence of what kind of treatment is most effective for a certain type of patients. Research data can provide some evidence of the effectiveness of treatments, however, in order to treat a specific patient, clinician still face the problem of how to choose the best therapeutic schedule by using existing evidence. New therapeutic schedule could be more effective than the conventional ones, but could be more expensive. Is the increase in the cost of treatment equivalent to the increase in the effect? Actually in medical practice, clinicians are usually keen to choose the more expensive treatment, sometimes even

choose without hesitating in the case of the lack of a reliable evaluation of the treatment.

The presence of the above-mentioned problems not only leads to the patients' health problems not being solved effectively, but also causes unlimited rise of health costs. As the Institute of Medicine (IOM) of the United States stated in 2008, "Patient care must be based on due diligence, clear and wise use of the best available evidence". It is to create and collect the best medical evidence, from which to implement the medical mission and to care for the patients as much as possible. The concept of Comparative Effectiveness Research (CER) has emerged in such context. In this chapter, we are going to introduce CER according to the report of the Congress of the United States in 2007 "Research on the comparative effectiveness of medical treatment: issues and options for an expanded federal role".

30.2 Definitions

According to the IOM of the United States, "Comparative effectiveness research (CER) is the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels".

CER is a sort of simple but rigorous assessments which mainly compares several alternative diagnosis/treatment methods possibly being chosen by certain population of patients. The compared methods can be analogous (e.g., medications with equivalent effects), or different (e.g., medication and surgery). Contents of CER include the comparison of advantages and risks of various types of diagnosis/treatment methods, as well as their cost effectiveness. Sometimes, a particular treatment could be proved to be effective or to have the highest cost effectiveness for the majority of patients, but the most critical and difficult thing is how to find the most effective and cost-effective method for a specific type of patient.

In many countries, the results on effect comparison are routinely used to determine the treatments covered by medical insurance and the proportion of Medicare reimbursement. For example, missions of the National

Institute for Health and Clinical Excellence (NICE) of the UK are to compare the clinical effects and cost-effectiveness of new and existing medications, methods and techniques, and to provide guidelines for the diagnosis/treatment for certain type of diseases or patients. By 2007, the NICE has published appraisals of over 100 specific technologies, guidance on the use of about 250 medical procedures, and about 60 sets of treatment Guidelines — a substantial but not exhaustive list.

The core of CER is to produce optimized evidence. To achieve the optimization of evidence, the following are needed. ('Initial National Priorities for Comparative Effectiveness Research'. Committee on Comparative Effectiveness Research Prioritization, Institute of Medicine)

(1) Study Populations Representative of Clinical Practice

Many studies on the effects of medical interventions on health address efficacy rather than effectiveness. Efficacy reflects the degree to which an intervention produces the expected result under carefully controlled conditions chosen to maximize the likelihood of effects. Many randomized controlled trials — generally considered to be the gold standard — are efficacy studies, particularly those conducted to win regulatory approval. The study population and setting of efficacy studies may differ in important aspects from those settings in which the interventions are likely to be used. By contrast, effectiveness research intends to measure the benefits and harms of an intervention in ordinary settings and broader populations, and therefore can often be more relevant to decision making and evaluation of health care providers and patients. However, since it is impossible to have randomization in the effectiveness researches such as observational, database, registry, and other studies, the unidentified bias and confounders may weaken the level of evidence; and the evidence may be strengthened by the efficacy studies in broaden populations or settings generating more generalized outcomes.

(2) Focus on the Individual Rather than the Average Patient

With the growing knowledge of disease mechanisms, systems biology, genomics, and other sciences that create the potential for more targeted therapies, doctors, patients and policy makers are increasingly seeking evidence not only from the general populations, but also from relevant subgroups. Increasing emphasis on patient-level attributes that may modify the balance

of benefits or harms can lead to more personalized medicine, reducing the pressure to try alternatives found to be ineffective in similar subgroups.

(3) Study Two or More Interventions by Direct Comparison

Although the public requires evidence of “active comparators”, comparing evidence-based alternatives (including usual care), and in some clinical circumstances, the government also support comparison studies comparing with “active control”, there is a paucity of head-to-head comparisons. Beyond specific medical interventions and technologies, there is a need for evidence evaluating the clinical and resource effects of innovations in health care delivery models, including new benefit designs, cost-sharing techniques, integrated organizational models, public health and population-level strategies, and interventions to improve the quality of care. Because these interventions are often implemented at provider or regional levels, the methods required to evaluate them may differ from those used to evaluate patient-level interventions.

30.3 Examples

The following are five examples of CER, which may help the readers to understand the methods and features of CER.

Example 30.1 “Meticulous analyses” may overthrow consensus on the comparative superiority of different therapies In patients with stable coronary artery disease, percutaneous coronary intervention (PCI) is the therapeutic method commonly performed by doctors. The study indicates that PCI reduces the incidence of myocardial infarction and death in patients who have acute coronary syndromes, but the result remains unclear for patients with stable coronary artery disease. Between 1999 and 2004, a randomized trial involving 2287 patients with stable coronary artery disease from 50 clinical centers was conducted in the US. One group was assigned to PCI with optimal medical therapy (PCI group, $n = 1149$) and the other group received optimal medical therapy alone (control group, $n = 1138$). After a follow-up period of 2.5 to 7.0 years (median = 4.6), although the PCI group showed better revascularization and less heart symptoms, there was no significant difference between the two groups in 5-year survival rate, incidence of myocardial infarction or stroke, and hospitalization rate

for acute coronary syndrome. In 2004, more than 1 million coronary stent procedures were performed in the US, and recent registry data indicated that approximately 85% of all PCI procedures were undertaken electively in patients with stable coronary artery disease. Therefore, the therapeutic method that generally accepted, by contrast, increases medical cost without substantial benefit (New England J. of Medicine (2007), 356, 1503–1516).

Example 30.2 Inexpensive “old drugs” are not necessarily inferior to new drugs Diuretic (e.g. Chlorthalidone) is a kind of inexpensive and effective antihypertensive drugs. With the development and usage of new drugs, antihypertensive drugs commonly used also include angiotensin-converting enzyme inhibitor (ACEI, e.g. Lisinopril) and calcium channel blocker (CCB, e.g. Amlodipine), etc. The optimal first-step therapy for the patients with hypertension, especially for those old ones with the risk of CDH, is unknown. Between 1994 to 2002, a total of 33,357 participants (aged 55 years or older with hypertension, and at least one other CHD risk factor) from 623 North American centers were selected for a randomized, double-blind, multi-center clinical trial conducted by the National Institute of Heart, Lung and Blood of the US. Participants were randomly assigned to receive Chlorthalidone ($n = 15,255$), Amlodipine ($n = 9048$), and Lisinopril ($n = 9054$) for planned mean follow-up of 4.9 years. As a result, there was no significant difference in the incidence of mortality, fatal CHD, or nonfatal myocardial infarction among the three groups. Systolic blood pressure of Diuretic group was lower than that of the other two groups. Diuretic has advantages on other indicators (e.g. stroke) as well. The results of comparison indicated that “old drugs” are more effective and cheaper, and should be selected as optimal first-step drugs (J. of American Medical Association, (2002) 288, 2981–2997).

Example 30.3 Surgery or non surgery? To discuss advantages and disadvantages by ‘head-to-head’ comparison Effects on patients with severe emphysema and criteria for the selection of patients have not been established. In a multi-center, randomized, controlled trial in the US, a total of 1218 patients with severe emphysema were randomly assigned to the lung-volume-reduction surgery group ($n = 608$) or the medical treatment group ($n = 610$). The result showed no significant differences in overall mortality between the two groups. After 24 months, exercise capacity improved

by 15% in patients of the surgery group, while by only 3% in patients of the medical-therapy group, and the difference was significant. By analyzing data of patients with predominantly upper-lobe emphysema and low exercise capacity, mortality was found lower in the surgery group than in the medical-therapy group. Overall, lung-volume-reduction surgery increases the chance of improved exercise capacity but does not confer a survival advantage over the medical therapy. Surgery confers a survival advantage for patients in some particular conditions. But for those with non-upper-lobe emphysema and high base-line exercise capacity, lung-volume-reduction surgery is not a wise choice. 'Head-to-head' comparison not only helps discerning the advantages and disadvantages of surgery, but also gives clues of indications for surgery (New England J. of Medicine (2003) 348, 2059–2073).

Example 30.4 For the manufacturer, comparison researches may be risky A pharmaceutical factory sponsored a randomized, double-blind trial. A total of 4162 patients who had been hospitalized for an acute coronary syndrome were randomly assigned into the pravastatin group ($n = 2063$, Drug A group, produced by the sponsor) and the atorvastatin group ($n = 2099$, Drug B Group), with a mean follow-up of 24 months. The results showed that the median of LDL cholesterol was 2.46 mmol/L in group A, and was 1.60 mmol/L in group B. The difference between the two groups was significant. By using Kaplan–Meier method, the outcome rates (death) after two years' follow up was 26.3% in group A and 22.4% in group B. The difference was significant. Although the sponsor hoped that drug A to be more effective, the study indicated that drug B worked better in preventing cardiovascular disease (CVD), death and lowering LDL cholesterol level among patients with acute coronary syndrome, as compared with drug A. As conducting comparison researches is risky, it is not easy for pharmaceutical factories to sponsor trials of this kind (New England J. of Medicine (2004) 350, 1495–1504).

Example 30.5 Is MRI effective for the examination of breast cancer? Comparison research can be used to discern the difference of efficacy between examination methods as well. A study in the Netherlands compared the efficacy of MRI with that of mammography for early-stage screening in women with a genetic or familial predisposition to breast cancer. They

found that the sensitivity of mammography and MRI for detecting invasive breast cancer was 33.3% and 79.5%, respectively, and the specificity was 95.0% and 89.8%, respectively. The difference was significant. Efficacy of MRI was obviously better than that of mammography, especially in patients with invasive tumors that were 10 mm or less. Another study conducted in Canada also indicated that the detection rate for breast cancer of MRI was higher than that of mammography, and no differences were found between the mortality rates of the two groups. The charge of MRI is high. Furthermore, although MRI had a higher detection rate of cancer, survival rate of patient was not increased by early detection. Issues on whether to promote MRI among women at high risk should be given careful consideration. (New England J. of Medicine (2004) 351, 427–437; J. of the American Medical Association, (2004) 292, 1317–1325).

30.4 Features and Principles

CER has six features as follows. ('Initial National Priorities for Comparative Effectiveness Research'. Committee on Comparative Effectiveness Research Prioritization, Institute of Medicine)

- (1) CER has the objective of directly informing a specific clinical decision from the individual patient perspective or a health policy decision from the population perspective. The range of potential objectives for CER studies gives the field a broad scope. Clinical questions refer to the health care of individual patients, including preventive, screening, diagnostic, therapeutic, monitoring, or rehabilitative interventions. Policy questions refer to the health and health care of populations through knowledge synthesis and transfer strategies, public health programs and so on. As CER contributes to such important decisions, all relevant stakeholders (including patients and the public) and decision makers would reasonably be included throughout the CER process, including priority setting, study design, explanation and implementation of results (Tunis *et al.*, 2003).
- (2) CER compares at least two alternative interventions, each with the potential to be "best practice". For many clinical decisions, "optimal usual care" is considered the standard which is used to be compared in CER. CER studies may also include the alternative of "watchful

waiting” when it is considered a reasonable strategy in the clinical context. CER highlights research that compares a test intervention with a method which is frequently-used and considered to be effective.

- (3) CER describes results at the population and subgroup levels. The primary outcome of a clinical trial is a measure of the “average effect” of an intervention, usually as estimated in the population assigned to the intervention in the trial. During the application of the results, clinicians must judge whether a particular patient is sufficiently similar to the trial population. By its focus on subgroup results, CER assists providers and patients in individualizing decisions — going beyond the average effects to the effect in subjects with common clinical characteristics.
- (4) CER measures outcomes — both benefits and harms — that are important to patients. There is an important distinction between much clinical research and CER, in that CER places high value on external validity, or the ability to generalize results to real-world decision making. Harms or risks of unintended consequences are also outcomes of interest, because they influence the net benefits of an intervention. Resource utilization may be highly relevant to net benefits when comparing the full clinical course of interventions, and cost-effectiveness analysis is a useful tool of CER.
- (5) CER employs methods and data sources appropriate for the decision of interest. CER includes at least three broad categories of research methods. Where evidence is lacking, CER may generate it either in non-experimental studies (observational settings) or in experiments (randomized controlled trial, as well as nonrandomized controlled trials). For decisions that have been the topic of substantial previous research, synthesis of existing studies (systematic reviews and meta-analysis and decision analysis) may be appropriate. Data sources for CER may thus include published studies, existing data from the delivery of care (insurance claims data and electronic health records), clinical registries, and information collected by clinical investigators, either retrospectively or prospectively.
- (6) CER is conducted in settings that are similar to those in which the intervention will be used in practice. For experimental studies, investigators should deliver the intervention in settings that are as close to actual practice as possible. Consistent with the definition of effectiveness, the

settings of CER studies are a defining characteristic. The settings of both experimental studies and observational studies are the realistic practice settings.

Although CER has the features described above, it does not mean that all CER should have the features. An intervention study can be compared with blank control group, placebo group or standard intervention group in the beginning. Actually, at the early stage of a new treatment study, safety and efficacy trials must be carried out in a specific setting. When the intervention is effective compared to placebo, the "head-to-head" comparison need to be carried out to answer the key question that which method is the most appropriate to the specific patient populations.

During the CER studies, researchers should follow the principles as follows:

- (1) The objective of CER is to help the decision-makers (patients, doctors, payers and policy makers) make wise decisions in health practices.
- (2) CER aims to discover and fill in the knowledge gaps in the practices.
- (3) CER provides the information to individual or population about the benefits, harms, fees and logic of different strategies and treatments.
- (4) The scope of the CER study is wide, including intervention, testing, prevention, health services, quality of service, etc.

Based on the principles above, CER must do the following:

- (1) The "head-to-head" comparison between several effective tests, treatments or prevention methods in the current standard must be carried out instead of comparing with placebo only.
- (2) To evaluate the outcomes which are directly related to the patients, instead of the benefits of science or one specific experimental indicator.
- (3) To compare the economics issues of different preventive and care measures, not only the effects but also the cost.
- (4) In order to help the patients and doctors make a choice among the effective treatments, the patients' characteristics which closely related to different outcomes must be identified instead of giving a general conclusion that the effect of one treatment is the best on average.

30.5 Research Methods and Techniques

A variety of methods are used to evaluate different treatment options, including the synthesis of existing research data (systematic review), or to carry out head-to-head clinical trials of different methods. Although these studies are not mutually exclusive, each category face certain challenges, and researchers need to weigh against the research costs and the value of the information obtained. With the increasing emphasis on CER, people are increasingly concerned about how to use the most appropriate method to study “head-to-head” medical problems (Congress of the United States in 2007 ‘Research on the comparative effectiveness of medical treatment: issues and options for an expanded federal role’).

30.5.1 *Systematic reviews of existing research*

Systematic review would probably be the easiest approach to compare effectiveness of different treatment options by reviewing and summarizing the results of existing research in a systematic and rigorous way. Many existing studies may only compare a single treatment to a placebo, but the results of several studies of individual therapies could in some cases be combined to measure those treatments against one another. This combination could critically evaluate the strengths and weaknesses of the existing evidence, to coordinate conflicting results, or to interpret the existing evidence. One advantage of the systematic review is its relatively low cost when compared to others.

As the evidence required to compare different treatment options is limited, how much additional insight can be gleaned from systematic reviews of existing research is not clear. Randomized controlled trials can provide the most conclusive evidence about the effect of the treatment among the existing evidences, so existing results from randomized controlled trials would naturally be the focus of any systematic review. But such studies also have limitations, one of which is that clinical trials sponsored by interested parties are more likely to get positive results than independent research. For example, after analyzing the data of 37 studies sponsored by commercial organizations, which is often the only source of such data, it was found that these studies were more likely to report results in favor of the sponsor (Mantel-Haenszel $OR = 3.60$, 95% $CI = 2.63-4.91$), suggesting that conflict of interests potentially impact the study.

Another potential limitation is that existing information may not be sufficient to reach a unified and clear conclusion. Studies may be difficult to compare or reconcile, either because they use different methodologies or analyze different populations of patients, or simply because they yield conflicting findings. For example, a number of independent studies have examined different screening techniques for colorectal cancer, each of which provides an estimate of the cost for each increase in quality adjusted life year (QALYs) per enrollee. But according to a recent review of those studies, the results varied to such an extent that reaching a clear conclusion about which technique was most effective or most cost-effective was difficult.

Comparative studies of available treatments may have even more limitations than studies of population screening tests, because trials of treatments for particular diseases usually exclude patients with certain health problems, elderly enrollees, or others who may be of considerable interest in gauging comparative effectiveness. As a result, it is hard to determine how broadly the results apply or whether they will hold for other groups of patients. The fundamental issue is that, no matter how rigorously a systematic review is conducted, its contribution is by definition constrained by quality of the underlying original evidence. For example, a systematic review of studies illustrates the advantages and disadvantages of diabetes therapy. The retrospective study covered a large body of literature, consisting of over 200 reports, and it was able to reach a relatively clear conclusion: Older drugs were found to be at least as effective as newer drugs in controlling patients' blood sugar and cholesterol levels. However, most of the studies that were reviewed had relatively short durations — two years or less — so they were not able to address the impact on mortality. At the same time, many studies have focused on the non-elderly white patients, and therefore cannot explain the therapeutic effect of these diabetes drugs on other ethnic groups. In addition, the study population of the research — diabetes patients excluded those with other serious health problems, but in practice, most diabetic patients suffering from other diseases at the same time, this distinction further limits the potential use of the research results.

Therefore, systematic reviews find that the available evidence is not adequate to address many important problems, so the primary value of such reviews may lie in identifying clearly the gaps in knowledge that should be the subject of future research. The statistical methods of systematic review can be found in Chap. 29.

30.5.2 *Randomized controlled trials*

Randomized controlled trials would probably be the research method that yields the most definitive results. But this approach would also be the most expensive and would take the longest to conduct (referred to Chap. 11 for more details). Rigorously designed trials will have an important impact on the selection of clinical programs. For example, in 2007, a study conducted in the US compared the efficacy of angioplasty and metal stent with nonsurgical management in the patients with stable coronary artery. It was found that the effect of metal stents failed to prevent death and myocardial infarction effectively. Although the researchers noted that the study results were “unexpected” and that its methods and results needed a careful discussion, the findings made a rapid reduction in the usage of stents (10% less than one month before the report was released, and 15% less as compared with the same period in the previous year). In another study, the researchers evaluated the effects of lung-volume-reduction surgery for advanced emphysema patients. The study found that several types of emphysema patients would benefit from the therapy and the health insurance agency (Medicare) agreed for nationwide coverage of the cost of treatment of these therapies, with the estimated expenses of approximately \$15 billion. But after the publication of this study, the actual number of patients underwent the procedure decreased rapidly, because the results showed that lung-volume-reduction surgery failed to extend the survival time with 10% mortality. The impact of the above randomized controlled trials for medical practice may not be typical, but publishing the results of these trials often takes several years.

Although the number of randomized controlled trials is increasing dramatically in recent years, there are still many problems. Many research evaluated efficacy rather than effectiveness. As mentioned earlier, the main difference between these two is that the former is a study under ideal conditions, while the latter is carried out in a real medical environment. In many randomized controlled trials, patients with other health problems or of certain groups (such as the elderly) are often excluded from the study. In addition, many of the study objects are patient who meet a certain definition, so its results may not be universal. At the same time, the objectivity of results of randomized controlled trials with commercial sponsorship is also being questioned.

Randomized controlled trial has some other limitations. First, the relatively high costs and long duration limit the feasibility of some effect studies. Secondly, the more stringent definition of the target population, the fewer patients are eligible to meet the inclusion criteria, and the higher the degree of difficulty of implementation. Again, if the pros and cons of different treatments cannot be confirmed, there will be no problem whatever groups the subjects are assigned to. However, if the subjects are assigned to a group which is generally considered to be less effective, there come the ethics problems. Given these limitations, carrying out and promoting comparative effectiveness research cannot rely solely on randomized controlled trials. It is necessary to combine randomized controlled trials with observational studies.

On the basis of the existing randomized controlled trials, a new method is proposed by using the computer model to simulate the therapeutic effects of treatments targeting at different patients. This method can serve as a substitution or complement of clinical trials. There are some mature models, and the most prominent may be Archimedes model designed by the team of David Eddy. The advantage of such method is once a model is built, the effectiveness of the particular method can be investigated with a relatively low cost. In fact, this approach can be even better than the analysis of claims data, electronic health records, or medical registration data. If the model can accurately predict the effects of a new treatment, the time waiting for those treatments to be used and tracking their effects on actual patients can be reduced. However, it may be quite difficult to achieve this goal. A prominent obstacle is that even if the model is rich enough to simulate the real situation of medical services, after all, it is not entirely true, and this makes it difficult to have enough confidence and acceptance of the results.

30.5.3 *Practical clinical trials*

To solve the existing problems in the randomized controlled trial, some researchers proposed that a greater emphasis should be put on "practical" clinical trials (practical clinical trials, pragmatic clinical trials, PCT) in comparative effectiveness study. Compared with RCT, there are three main features of PCT.

(1) Comparing alternative methods that clinicians face in practice

Example 30.6 A PCT conducted by the Department of Veterans Affairs (VA) compared the effects of terazosin hydrochloride (TH) and finasteride (FI) in patients with benign prostatic hyperplasia (BPH). Both medications have been approved by the Food and Drug Administration (FDA), but the evidence was based the comparisons with placebo, and the manufacturers had no motivation to initiate comparative studies of the two medications. In this PCT, 1229 BPH patients were randomly assigned to placebo, FI, TH or FI-TH combination groups. The average changes of symptom scores in the first year since baseline was 2.6, 3.2, 6.1, and 6.2 for the four groups, respectively, and the FI, TH and FI-TH groups all had significant differences as compared to the placebo group ($P < 0.001$). The increase of peak urinary-flow rates in the first year was 1.4, 1.6, 2.7, and 3.2 ml per second for the four groups, and comparing TH and FI-TH groups with FI and placebo groups, the differences were all significant ($P < 0.001$). In conclusion, TH was more effective than FI (once the study initiated, both manufacturers contributed to the design and funding for the study) (New England Journal of Medicine (1996) 335, 533–9).

(2) Extracting a wide variety of participants from clinical patients

RCTs are generally more stringent in the inclusion of participants, while PCTs have relatively broad inclusion criteria, and the exclusion criteria are also as lenient as possible. Patients recruited in a RCT are usually with clear diagnosis, but in practice, doctors often need to start symptomatic treatment before diagnosis, therefore expanding the participants to patients with certain symptoms is much meaningful.

Example 30.7 Before the treatment of patients with rhinosinusitis, doctors usually judge the patients' condition by symptoms and results of X-rays, rather than conduct sinus puncture immediately to diagnose. In a multicenter trial of sinusitis treatment, the inclusion criteria required the patients having sinusitis symptoms or being found sinus infections by X-rays, but not patients diagnosed by sinus puncture. Patients were allocated to the intervention group (nasal corticosteroid therapy combined with conventional medication) or the control group (placebo nasal spray combined with conventional medication) blindly and randomly. The study found that

nasal corticosteroids combined with conventional medication therapy can improve treatment effects as compared to the control group. Because the study expanded the subjects from those who were limited and were difficult to be diagnosed in community hospitals to those who were broader and suitable for the operation in community, the results have more practical applications for community physicians. In addition, PCTs often recruit subjects through community and primary hospitals, hence the patients are more diverse and the findings could be more useful for general medical practice (JAMA (2001) 286, 3097–3105).

(3) Collecting a wider range of health-related outcomes

In addition to traditional research outcomes (e.g., death and morbidity), results of PCT results also include indicators related to function (e.g., quality of life, severity of symptoms, satisfaction, costs of treatment, etc.). Meanwhile, in order to reflect the natural history of diseases as much as possible, the follow-up period of PCT is usually longer than that of traditional clinical trials. The results obtained from the two types of studies may hence different. For example, in a study of surgical repair of abdominal aortic aneurysm, most of the previous studies had short follow-up period, so had left substantial uncertainty. In two other studies, which were followed up for 4.9 and 8.0 years, results showed that the survival rate was not improved by elective surgical repair of aneurysms.

PCT is simpler, less expensive and less time cost, so is regarded to be more “worthwhile”. But it also faces the risk of reduced accuracy.

Example 30.8 In a study of the effects of hormone treatment for the menopause, 16,608 healthy menopausal women were randomly assigned to the hormone therapy group or the placebo group, with an expected follow-up period of 8.5 years. After an average follow-up period of 5.2 years, results showed that the incidence of invasive breast cancer in the hormone therapy group exceeded the preset stopping boundary, and the risk of cardiovascular diseases was high, so the study was stop immediately. However, further analysis found that the effect of hormone therapy was related to the patient’s age, and hormone therapy was beneficial for some patients. (JAMA (2002) 288, 321–33).

30.5.4 Cluster randomized trial

Interventions of RCT is generally targeting at individuals, but some studies need to adopt cluster randomized trials (CRT), especially when the intervention is to be applied to an entire group (e.g., a community-based health promotion initiative), or when the status of individuals are linked (e.g., studies of contagious diseases). In the study carried out at methadone maintenance treatment clinics aiming at reducing dropout rate, the participants were randomly assigned to the intervention group (receiving psychological and behavioral counseling) and the control group (receiving conventional treatment), the interaction between the participants may affect the control group (contamination), hence may affect the judgment effects of the intervention. Comparison between CRTs and RCTs is showed in Fig. 30.1.

In CRTs, unit of randomization could be community (e.g., to carry out mass media education), clinic (e.g., to carry out medical intervention), school (e.g., to carry out the smoking prevention intervention) or family (e.g., to carry out diet intervention). There are several advantages of CRT.

- (1) By applying interventions at the hospital or community level, CRT can more readily study interventions under conditions of actual use. For instance, a CRT that uses existing clinical and administrative mechanisms incorporates the impact of group dynamics (advocacy, peer pressure, reminders) among healthcare providers.

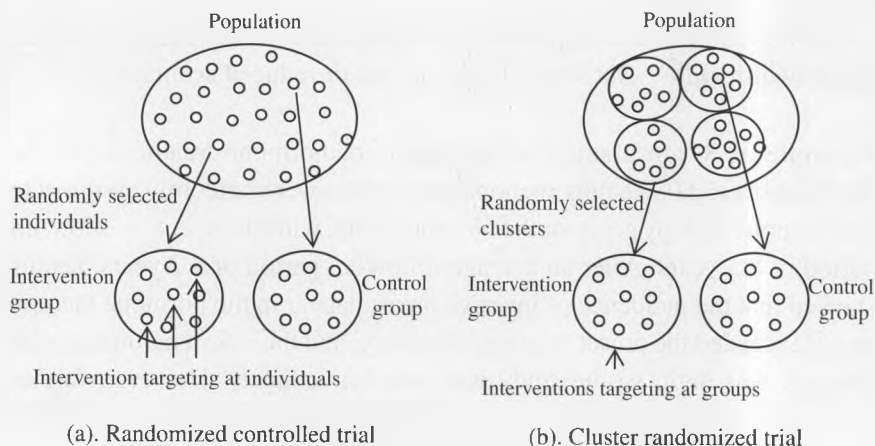


Fig. 30.1 Comparison between randomized controlled trial and cluster randomized trial.

- (2) CRTs are often intended to be applied to an entire hospital, Intensive Care Unit (ICU) or clinic population without exclusion, which enhances generalizability.
- (3) CRTs are able to harness the health care delivery system's existing administrative capacities, including quality improvement programs and data collection systems, simplifying the logistics of implementation and reducing study costs. The increasing availability of electronic health information facilitates the implementation of cluster randomized trials, as routinely collected electronic health information can be used to assess baseline status, monitor implementation, and measure outcomes.

Example 30.9 In a study of the prevention of the Methicillin-resistant *Staphylococcus aureus* (MRSA) infection in ICU, the researchers used CRT design to compare the effect of three MRSA control strategies (active screening and isolation, active screening and MRSA eliminate colonization, widely cancel colonization without considering the status of MRSA). All the three control strategies are used in practice, but which is better is inconclusive. Affiliated hospitals of the Hospital Corporation of America (HCA) ranked the hospitals according to the annual number of ICU patients, and then divided the hospitals into six groups according to the rank (six groups with six hospitals, and a group with three hospitals). Hospitals within each group were further ranked in accordance with MRSA prevalence, the adjacent three hospitals were randomly assigned to three MRSA control strategies, and ICU in each hospital received the same control strategy. This randomized method balanced the number of ICU patients and MRSA prevalence in each group. The main outcome of the study is the number of patients with the infection of MRSA during two days after ICU admission to two days after leaving the ICU. Other outcomes include MRSA infection in the blood or urinary tract. All results, including the cost of treatment and so on, can be obtained from the information system of HCA. The study used existing vocational education systems and compliance monitoring systems. The executors and the supervisors are the person in charge of the ICU or infection control management, rather than specially trained researchers. This kind of research is based on existing systems and staff, making the intervention and evaluation closer to reality, and is conducive for the information exchange within the HCA system.

What to be noted in CRT is that the research often does not require informed consent of each subject, so there may be ethical issues. CRT design must also meet ethical requirements. In addition, due to the non-independence of individuals within the same cluster, the effective sample size is less than the total number of individuals. This reduction of the sample size depends on the average group size (m) and the degree of correlation (ρ) within the group. In order to compensate for reduced power of the cluster randomized controlled trials, the sample size should be expanded on the basis of the sample size of individual randomized controlled trials to $[1 + (m - 1)\rho]$ times.

30.5.5 *Observational study*

Compared with randomized controlled trials, pragmatic clinical trials and cluster randomization trials, observational studies do not assign interventions for the exposure of objects. It is one type of methods used to describe the distribution of diseases and health among population, and to explore the sequential relationship between exposures and diseases, through field test, analysis and recording data objectively. Traditional observational studies include cross-sectional study, case-control study and cohort study. It is refreshed with the development of registration data and claims records. Methods based on such data have been receiving increased attention.

Considering its low internal validation and confounders, observational studies are usually characterized as having inferior quality of evidence than randomized controlled trials. While more and more researches indicate recently that well-designed observational studies can provide information about intervention effect effectively, and the quality of evidence is as good as randomized controlled trials. In contrast, some comparative effectiveness questions are particularly appropriate to be answered by observational studies:

(1) When large studies are needed

Example 30.10 Although multiple treatment guidelines recommend the use of systemic corticosteroids for flare-ups of chronic obstructive pulmonary disease, the optimal dose and route of administration have not been well defined. Data from an observational study of 84,621 patients conducted at 414 US hospitals in 2006 and 2007 demonstrated that physicians

were more likely to administer high-dose (average 600 mg/d) intravenous systemic corticosteroids in such cases rather than low-dose oral corticosteroids (average 60 mg/d). Of all patients, 92% were initially treated with intravenous systemic corticosteroids, whereas 8% received oral corticosteroids. The death rates and adverse event frequencies were 1.4% and 10.9% in the intravenously treated patients while 1.0% and 10.3% in orally treated patients respectively. When confounders were adjusted by propensity-matched analysis (more detail of the method are introduced in “useful methods”), the treatment effects of the two therapies were in fact quite similar, while the risk of side effect was lower among orally treated patients, as was length of stay and cost (JAMA (2010) 303, 2359–67). To evaluate these findings in a randomized trial, the trial would have to be very large, with approximately 30,000 patients in the two groups combined (JAMA (2010) 303, 2409–10).

(2) When treatment adherence differs

Example 30.11 Patients usually have poor compliance in the use of inhaled steroids, which are considered to be the gold standard for treating asthma. An insurer requested observational study conducted in US using administrative claims data of about 51,168 patients who have at least one record in September 2003 and August 2005, the study evaluated the clinical, economical and treatment effect of patient to explore the association between asthma medications and outcomes. The researchers concluded that although inhaled corticosteroids were associated with a lower risk of inpatient admissions and emergency department visits, patients taking oral medications were significantly more likely to adhere to their treatment regimen. Even when the investigators controlled the severity of disease, patients taking oral medications have greater benefit from treatment than those taking inhaled medications. Based on the study, the insurance company decided to continue its favorable reimbursement level for the oral medication. However, treatment adherence differs will influence the outcomes in randomized trials if it is unable to realize blindly. (Mayo Clinic Proceedings (2009) 84, 675–84).

(3) When providers have different trainings

Implantable cardioverter defibrillators for patients at risk for sudden cardiac death can be implanted by physicians with a range of training, from

accredited electrophysiology fellowships to less formal training programs. A study of physician certification and outcomes in 111,293 patients showed that patients had a higher rate of procedural complications when the devices were implanted by thoracic surgeons or cardiologists who were not electrophysiologists, compared to when the devices were implanted by electrophysiologists. Furthermore, among 35,841 patients who met the standard criteria for use of defibrillators with cardiac resynchronization therapy, patients were more likely to receive the indicated resynchronization device when their defibrillators were implanted by an electrophysiologist. That is, electrophysiologists were more likely to implant the appropriate type of device than those who were non-electrophysiologist cardiologists. In randomized trials, once the providers themselves will influence the outcomes, they must be balanced between groups during the design stage, while such method can make the design more difficult, and even infeasible in some clinical practices.

As the confounding cannot be balanced through randomization in observational studies, such type of researches is more prone to bias. Clear and transparent reports are helpful for readers to evaluate potential bias and confounding in a certain research. In order to regulate the report quality of observational study, an international cooperation group constitutes of epidemiologists, statisticians, famous magazine editors and clinical doctors have made great efforts to develop the reporting standard of epidemiological observational study — Strengthening the Reporting of Observational Studies in Epidemiology (STROBE). More details can be found on the website (<http://www.strobe-statement.org>). But the STROBE is mainly aimed at cross-sectional study, case-control study and cohort study. Recently, with the development of comparative effectiveness studies, researchers put forward good research for comparative effectiveness (GRACE) principles to guide on how to design and evaluate observational comparative effectiveness studies. The GRACE principles include three aspects: first, specified study plans (including research questions, main comparisons, outcomes, etc.) are needed before conducting the research; second, the study should be conducted and analyzed in a manner consistent with good practice and reported sufficient detail for evaluation and replication; third, the interpretation of comparative effectiveness for the population of interest derived from the research should be valid.

30.5.6 *Analyses of claims records*

A somewhat more challenging approach than others would be health insurance claims records. An advantage of that approach is that it could provide new information to help resolve uncertainties about treatments at relatively low cost — using data on patients that had already been treated.

A central difficulty in such studies, however, is accounting for the differences in patients' health status that play a role in determining which treatment they get — which can make simple comparisons misleading. Insurance claims typically do not include any information about health status. Yet patients with more severe heart disease, for example, are more likely to receive invasive and expensive surgical procedures such as an angioplasty or a bypass operation. The greater severity of their condition may also make them more likely to have a subsequent heart attack and more likely to die. As a result, a comparison with patients receiving less aggressive treatments — who are probably not as sick, on average, to begin with — could understate the benefits of more aggressive treatments.

Other issues surround the claims data themselves. First, maintaining the privacy of the patients whose records were being examined would be an important matter but could also present a barrier to conducting such studies. Second, in order to obtain sufficient statistical power and significant findings, a large amount of claims data would be needed. Third, the quality of the study that could be conducted would depend on the level of detail that the data provided. Comparisons of the effects of treatments on mortality rates would be easier to generate because that information is relatively easy to obtain. Effects on morbidity or on the extent, to which symptoms are relieved, however, might be more difficult to ascertain — depending on whether the relevant data were readily available. In addition, private health plans might have difficulty in conducting longer-term comparative effectiveness studies using claims data on their enrollees given the turnover in insurance coverage; if patients who changed plans were different from those who remained, statistical obstacles might undermine the comparison.

30.5.7 *Medical registries*

Another option that could supplement or help improve analyses of claims data would be to establish medical registries, which generally track patients

who have a particular disease or who have received a specific treatment. Registries collect additional information that is typically not contained in claims records, such as measures of health status or test results. In the United States, a number of registries — established or managed by various entities, including medical specialty societies and product manufacturers — have been used to help determine the clinical effectiveness or cost-effectiveness of various products and services. Some health plans establish registries of their enrollees, although a centrally managed registry would have the advantage of being able to track patients if they moved or changed health plans.

Data from medical registries could help improve claims based analyses both by allowing a broader set of outcomes to be measured and by providing information to control for differences among patients getting different treatments, including the severity of their illness. But a number of challenges and trade-offs would exist. One issue would be how to recruit patients and their providers to participate in and provide information to the registries and to retain them over time. Voluntary participation might be easy to implement, but could introduce bias into analyses if the patients choosing to participate differed in important ways from the patients who had opted out. Some form of mandatory participation could avoid that problem, but might raise objections from participants. Registries focused on specific treatments could also be subject to bias if those patients differed systematically from patients who did not receive those treatments — a problem that could be addressed by including a comparison group in the registries. Another trade-off concerns the data elements to be collected; a more extensive list would permit richer analyses but would raise the burden of participation. More-extensive registries and registries involving more patients would also be more expensive to operate, although the annual costs of maintaining a typical registry are probably on the order of several million dollars.

The establishment of registries could affect medical practice in various ways. For example, Centers for Medicare & Medicaid Services (CMS) recently instituted a policy of “coverage with evidence development” for Medicare, to address treatments with potentially promising but uncertain medical benefits. Under that policy, Medicare now covers the costs of implantable cardioverter-defibrillators for a broader set of heart conditions than had previously been eligible — but only if those new patients are included in a registry that is supposed to track their progress. The new

policy allows broader access to that technology in order to help generate the kind of evidence needed to reach a conclusion about its value. Another example comes from Sweden, where a registry of patients undergoing hip replacement surgery has been used to provide periodic feedback to doctors about their surgical techniques and to track which specific models of artificial hip have the lowest rates of complications. That effort is credited with reducing health costs by avoiding repeat operations to fix faulty or poorly installed hips.

30.5.8 Useful techniques

30.5.8.1 Instrumental variables

In the analysis based on observational data or health insurance claims records, a potential risk is selection bias. For instance, in the effect research among different treatments for heart disease patients, those who have more severe heart disease are more likely to receive invasive and expensive surgical procedures such as an angioplasty or a bypass operation. The greater severity of their condition may also make them more likely to have a subsequent heart attack and more likely to die. As a result, a comparison with patients receiving more conservative treatments — who are probably not as sick, on average — could understate the benefits of more aggressive treatments. In other settings, patients receiving more aggressive treatments may be healthier, so even well-designed observational studies can generate misleading findings regarding the benefits of these treatments.

To address such problems, corrected methods can be used in statistics to control the confounder influences between groups, but if there exist unknown confounders, other methods are needed. Instrumental variables are commonly used in economics and now spread to medical field to analyze the relation between treatment and health effect. By observation we can obtain instrumental variables which are correlated with the treatment that patients receive but are not correlated with their underlying health. As illustrated in Fig. 30.2, by using instrumental variables we can simulate to randomly allocate patients into different treatment groups.

For example, one study using claims data from health insurance sought to explore the influence of intensive treatment (such as an angioplasty or a bypass operation) on elderly acute myocardial infarction patients. The data showed patients living farther away from hospitals were less likely

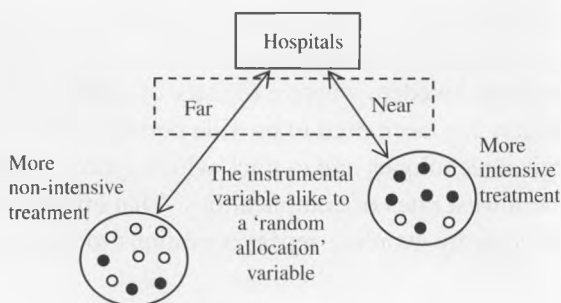


Fig. 30.2 Instrumental variables illustration.

to receive an intensive treatment while there was no difference on health condition between patients living farther and living nearer. Under this circumstance (living distance was not correlated with their underlying health but was correlated with the treatment that patients received), living distance could be used as an instrumental variable: patients living farther away were more likely to receive non-intensive treatment (as randomly allocated to non-intensive treatment group) while those living nearer were more likely to receive intensive treatment (as randomly allocated to intensive treatment group). The study found that patients receiving intensive treatment (living nearer) had slightly lower mortality rates, but the difference arose only on the first day of admission. In the long run, the study implied that intensive treatment of elderly acute myocardial infarction patients had no influence on mortality. But in the claims data analysis, all the studies do not have proper instrument variables. Even though there might exist such variables, the influence from other confounders on the result cannot be excluded easily.

30.5.8.2 Propensity score

With the development of social informatization, medical registries are substantially accumulated and continuously improved. Issues on how to reasonably use these observational data to compare the effects of different treatments and interventions deserve exploration. In the analyses of medical registries, it is difficult to demonstrate the real effects of interventions because of the unbalance of basic information and severity of diseases between different groups. Common ways to deal with the potential

confounders during data analyses include stratified analysis and adjusted analysis. In recent years, researchers proposed the application of propensity score to balance the influences of confounders between different groups.

Propensity score is a value related to the probability ratio of one choosing to use intervention (or the treatment being studied) instead of control when observable covariates are given. Propensity score could be calculated with logistic regression. When the dependent variable $Y = 1$ refers to choosing the intervention (or treatment) and $Y = 0$ otherwise, and the independent variables X_1, X_2, \dots, X_k refer to the potential confounders, the logistic equation for $P = \Pr(Y = 1 | X_1, X_2, \dots, X_k)$ can be built as follows:

$$\ln \left(\frac{P}{1-P} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k.$$

The propensity score for the i th subject is defined with

$$PS_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k.$$

One can see that the propensity score is a composite score of important confounders which well reflects the propensity of an individual being subject to the group of intervention (or treatment).

Methods of balancing confounders between groups with propensity score are as follows:

- (1) Adjustment. It is the method of directly including propensity score in the model as an independent variable and analyzing the relationship between the intervention and the outcome after adjusting propensity score.
- (2) Stratification. It is the method of dividing subjects into several strata and analyzing the relationship between the intervention and the outcome within each stratum.
- (3) Matching. In this method, all the subjects in different groups are respectively ranked according to their propensity scores. Each subject of the intervention group could be matched up with one or more subjects (randomly selected if there are multiple eligible ones) from the control group by similar propensity score.

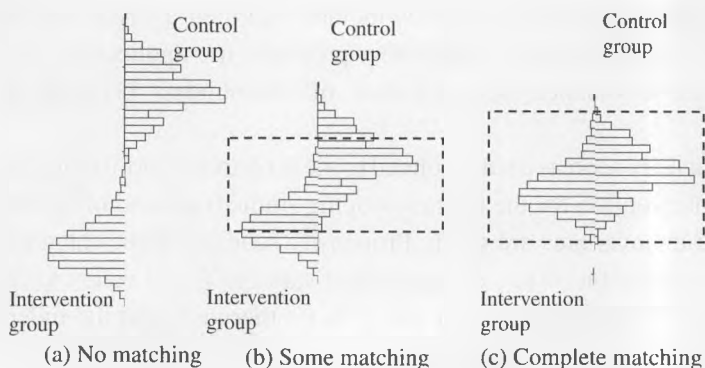


Fig. 30.3 Different matching status of propensity scores.

The newly constructed data will then be used to analyze the relationship between the intervention and the outcome of the study by using pairwise-designed methods. Of course, some subjects may be excluded from analysis because of the absence of suitable matching. Thus the procedure of matching may lose part of information. Three different kinds of matching are depicted in Fig. 30.3. In situation A, there is no overlap in propensity scores between the intervention group and the control group, which indicates failure in matching. In situation B, part of the members in two groups with their propensity scores in the overlapped range could hopefully be matched. In situation C, the propensity scores of the two groups are completely overlapped, and the matching could be performed to the best.

In a research conducted in Denmark (European Heart J. (2011) 32, 1900–8), multiple medical registration data and propensity score were used to compare the mortality and CVD risk of different insulin secretagogues (ISs) and metformin in type II diabetes patients with or without previous MI. Because of huge cost and great variety of insulin secretagogues, it is unrealistic to conduct a large-sample RCT. Every resident in Denmark has a unique, permanent registration number which is connected to various domestic registrations. Registration data used in the study included “The National Prescription Registry of Danish” (record information of all medication prescriptions since 1995), “The National Patient Registry” (record all the primary diagnoses of hospitalizations at discharge since 1978), and “The National Causes of Death Register” (record the information about causes of death). All individuals aged 20 years and above who initiated single-agent

treatment with an IS or metformin between 1997 and 2006 were included in the study, while those treated with insulin alone or with the combination of multiple drugs were excluded. A total of 107,806 patients with type II diabetes were included, and were followed up for up to 9 years (median 3.3 years).

Results of data stratified according to the presence of MI indicated the correlation between the usage of insulin secretagogues (glimepiride, glibenclamide, glipizide, gliclazide, tolbutamide and repaglinide) or metformin and the outcomes (all-cause mortality, cardiovascular and mortality). But the basic information (age, gender, treatment duration, etc.), comorbidity and drug combination were incomparable (partly listed in Table 30.1). These factors potentially confounded the correlation between the drug and the outcomes. Therefore, a logistic equation was built for treatments (a certain insulin secretagogues or metformin) according to potential confounders in the baseline, propensity score was then calculated for each subject, and a new sample was obtained after screening individuals by matching. Analyses of matched data indicated a better balance (partly listed in Table 30.2) in baseline characteristics (basic information, comorbidity, drug combination, etc.) between matched groups. Then adjusting the factors remaining unbalanced after matching, Cox regression analysis further indicated the inferiority of six insulin secretagogues to metformin in prevention of all-cause mortality and cardiovascular death. The difference of risk of all-cause mortality and cardiovascular death between glimepiride, glibenclamide, glipizide, tolbutamide and metformin was statistically significant respectively.

30.6 Steps of CER

There are seven primary steps in the implementation of CER:

- (1) Identify new and developing clinical interventions. To find existing problems by clinical practice and propose new clinical methods can be regarded as the object of comparative effectiveness research. These methods include all those related to patients' health in the process of disease prevention, diagnosis and treatment.
- (2) Review and summarize current medical research. Focusing on the medical issues being studied, widely review the related literatures and get

Table 30.1 Comparison of basic information and comorbidity of type II diabetes patients (without MI) who accepted treatment of metformin or insulin secretagogues.

	Metformin	Glimepiride	Gliclazide	Glibenclamide	Glipizide	Tolbutamide	Repaglinide
<i>N</i> (%)	43,340 (54.3)	36,313 (37.0)	5926 (6.0)	12,495 (12.7)	6965 (6.1)	5335 (5.4)	2513 (2.6)
Age (years)	52.5 ± 14.0	60.9 ± 13.3	60.0 ± 13.2	63.2 ± 13.7	63.0 ± 13.5	64.4 ± 13.5	57.9 ± 12.6
Men (%)	50.9	55.3	56.5	54.4	54.1	53.8	56.0
Treatment duration (year)	1.76 ± 1.58	2.11 ± 1.75	2.10 ± 1.75	2.35 ± 2.08	2.35 ± 2.08	2.36 ± 2.13	1.97 ± 1.76
Congestive heart failure (%)	1.1	2.5	1.6	2.4	2.4	2.6	0.7
Cardiac dysrhythmia (%)	1.6	3.2	2.1	3.0	3.2	2.8	1.5
Peripheral vascular disease (%)	0.3	0.6	0.5	0.7	0.9	0.9	0.6
Cerebrovascular disease (%)	1.6	2.8	1.4	2.9	2.8	3.3	1.2
Chronic pulmonary disease (%)	1.5	2.6	1.6	2.4	2.8	2.6	1.2

Data sources: Schramm *et al.* (European Heart J. (2011) 32, 1900–8).

Table 30.2 Comparison of basic information and comorbidity of type 2 diabetes patients (without MI) who accepted treatment of metformin or insulin secretagogues after propensity score calculation and matching (part of medication groups).

	Metformin	Glimepiride	Metformin	Gliclazide	Netformin	Glibenclamide	Metformin	Glipizide
<i>N</i> (%)	22 340 (50.0)	22 340 (50.0)	4 739 (50.0)	4 739 (50.0)	7 412 (50.0)	7 412 (50.0)	4 981 (50.0)	4 981 (50.0)
Age (years)	57.1 ± 12.0	57.3 ± 12.2	60.0 ± 13.2	60.0 ± 13.2	59.6 ± 12.9	59.6 ± 13.0	61.5 ± 12.7	61.6 ± 12.8
Men (%)	55.1	55.7	56.3	56.8	54.9	54.9	55.4	55.4
Treatment duration (year)	2.1 ± 1.7	2.1 ± 1.8	2.0 ± 1.8	2.1 ± 1.9	2.5 ± 2.1	2.4 ± 2.1	2.4 ± 2.1	2.5 ± 2.2
Congestive heart failure (%)	0.7	0.7	0.2	0.2	0.6	0.6	1.2	1.7
Cardiac dysrhythmia (%)	1.1	1.1	2.2	1.9	0.9	0.9	1.4	1.4
Peripheral vascular disease (%)	0.4	0.1	0.3	0.5	0.1	0.1	0.2	0.2
Cerebrovascular disease (%)	1.5	1.5	0.5	0.5	1.2	1.2	1.2	1.2
Chronic pulmonary disease (%)	1.1	1.1	0.7	0.7	1.5	1.7	1.4	1.4

Data sources: Schramm *et al.* (European Heart J. (2011) 32, 1900–8.

complete understanding of the current situation, existing methods and their shortcomings, so as to prepare for the design and development of a comparative effective research.

- (3) Identify the gaps between existing medical research and the needs of clinical practice. Identify problems in clinical practice that cannot be answered by existing data and design scientific research program.
- (4) Promote and generate new scientific evidence and analytic tools. From the results of comparative effective research, generate new scientific evidence or increase the completeness and accuracy of original evidence. At the same time, develop new methods to adapt to the needs of comparative effectiveness research.
- (5) Train and develop clinical researchers. In the process of comparative effectiveness research, pay attention to strengthen the scientific research ability of clinical staffs. Encourage them to discover and put forward questions from clinical practice; and unite the researchers from multiple disciplines to cooperate in the comparative effectiveness research.
- (6) Translate and disseminate research findings to diverse stakeholders. On one hand, the results of comparative effectiveness research should be used to guide clinical practice, which is the fundamental goal of CER; on the other hand, the achievements should be disseminated timely and widely to other clinical staffs by means of paper, report etc.
- (7) Inform stakeholders via an open forum. Beside clinicians, CER findings should be known by stakeholders in related areas like public health, health policy etc. Only in this way, can the studied issue be emphasized and promoted on every layer of stair.

30.7 Standards for Implementation and Report

Implementation and report of CER should comply with the following standards:

- (1) Theme and researchers. Both the patients and decision makers should join in the theme choosing and refining of CER. The team for CER must represent clinical or public health practice. As CER is aimed at solving specific problems in medical practice, the content studied must has direct relation to the real health problems faced with patients. It is necessary for the CER researchers to have the most advanced knowledge

as well as medical practice level to make CER by existing strategy and then generate new medical evidence.

- (2) Protocol. The protocol of CER should have high quality and transparency. Keep up to the highest scientific standards in the aspects of program design, data analysis and results interpretation, especially to the guidelines aimed at improving clinical research quality and transparency such as Consolidated Standard of Reporting Trials (CONSORT) for randomized control test and Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) for observational study. There should be explicit, clear and executive research program, focused issue, methods and analysis plan for a CER. Researchers should abide by the program strictly. Once it happens to some adjustment, all the changed should be recorded in detail. The programs of CER should have open access and be available for both researchers and patients by registration on the international institution websites.
- (3) Peer review. Strict peer review is needed for a CER report. Before publication, the research results must be reviewed by independent experts in this field, methodologists and statisticians. Based on the query proposed by experts, researchers need to revise and refine the report. In the meanwhile, the report should conclude a discussion about its limits such as bias, confounders and application range to make the report as fair as possible. In addition, to guarantee the public and other researchers to acquire the results without obstacles, the relevant magazines and media should provide CER results freely.
- (4) Compelling policy for conflicts of interests. Due to the potential interest influence on the assessed intervention, CER must comply with policies relating to interests unconditionally. For example, a certain CER finds some treatment is superior to the conventional treatment, which may lead to increasing usage of this treatment and decreasing usage of the conventional treatment during the subsequent clinical practice. So in the peer review and publication in any situation, the researchers, sponsors and other contributors must announce all the interest relationship clearly and leave the readers alone to judge the fairness of CER results.
- (5) Full cooperation with statisticians and epidemiologists. For statisticians and epidemiologists, CER is both of opportunity and challenge. On the majority, it belongs to observational study, but it is not completely equal

to traditional cohort study, case-control study and cross-sectional study. In order to make the CER fit to the high-level scientific and ethical standards, statisticians should not only use the existing technology, but also develop new methods on design and analysis to meet the specific demand of CER. At the same time, the variety of CER determines that statistics and epidemiology will encounter new challenges inevitably, such as adroit design of various observational clinical researches, management of severe missing data, imperfect follow-up, immeasurable bias, potential effect of chance and data mining techniques. Only by full cooperation with statisticians in the process of design, implementation, analysis and paper writing, can clinical researchers achieve higher level in CER.

- (6) The responsibility of medical magazines: medical magazines should advocate, develop CER and at the same time promote its research achievement. Magazines and peer reviewers must make sure that CER satisfies the highest scientific and ethic standards as well as other health-related research. So they need to develop methodology and statistics to evaluate the methods used in new or unfamiliar health care research properly.

30.8 Summary

- (1) CER is a brand new idea in medical research, which provides evidence for patients, doctors and administrators to make wise decision by comparing different methods in prevention, diagnosis and treatment.
- (2) CER is a kind of medical research based on “real word”, which combines medical practice to make “head-to-head” comparison. It concludes studied not only on medical effect and cost-effect, but also optimal studies on different subjects, different illness conditions and different aims.
- (3) CER needs large sample and it must be combined with information technology, data mining technique and informatics in medical field.
- (4) Magazine editors and statistical epidemiologists must correct their perspectives on CER and impel the clinicians to carry out CER during the process of clinical practice.

30.9 Computerized Experiments

Experiment 30.1 Propensity score matching To study the impact of two surgical methods *A* and *B* on survival time of patients with lung cancer, a total of 902 patients participated in the study, of which 562 were treated with *A* (63.3%) and 340 were treated with *B* (37.7%). In clinical practice, the doctor usually select appropriate surgical approach according to the actual situation of the patient instead of a random, thus, the baseline covariates between the groups may be distributed unevenly. To reduce such bias when comparing survival time of the two groups, matching by propensity score was used in this study to balance the baseline covariates between the two groups, and matched data were analyzed by survival analysis.

Step 1 Calculation of propensity score: Taking surgical method (variable VATS2) as the dependent variable ($Y = 0$ for *A*, and $Y = 1$ for *B*), and gender (gender), age (age), pathological characteristic (path), tumor stage (latesTNM) as independent variables to build a logistic regression model. The propensity scores can be calculated for each patient (Program 30.1), which is the probability of the patient receiving surgery *B* given the values of existing covariates.

Lines 01 to 04 of Program 30.1 input the data to SAS database; lines 05 to 11 are logistic regression taking VATS2 as the dependent variable and

Program 30.1 Calculation the propensity score.

Line	Program
01	PROC IMPORT OUT=A;
02	DATAFILE="H:\ Experiment 30-1.XLS";
03	DBMS=EXCEL REPLACE;
04	GETNAMES=yes;
05	DATA CO;
06	SET A;
07	PROC LOGISTIC DATA = CO;
08	CLASS path latesTNM ;
09	MODEL VATS2= gender age path latesTNM ;
10	OUTPUT OUT=co PROB=prob ;
11	RUN;

Table 30.3 Balance comparison of covariates of the two groups before and after matching.

	Before matching			After matching		
	Group A	Group B	P-value	Group A	Group B	P-value
Sample size (%)	562 (62.3)	340 (37.7)		259 (50)	259 (50)	0.671
Age Mean (\pm S.D.)	57.62 \pm 10.35	60.71 \pm 10.51	<0.001	59.70 \pm 9.65	59.32 \pm 10.70	
Gender (%)						
Male	419 (68.0)	197 (32.0)	<0.001	177 (50.7)	172 (49.3)	0.639
Female	143 (50.0)	143 (50.0)		82 (48.5)	87 (51.5)	
Pathological type (%)						
1	177 (69.1)	79 (30.9)	0.002	76 (50.7)	74 (49.3)	0.778
2	275 (57.1)	207 (42.9)		143 (50.7)	139 (49.3)	
3	110 (67.1)	54 (32.9)		40 (46.5)	46 (53.5)	
Tumor stage (%)						
1	190 (47.1)	213 (52.9)	<0.001	138 (49.8)	139 (50.2)	0.901
2	136 (71.6)	54 (28.4)		45 (48.4)	48 (51.6)	
3	236 (76.4)	73 (23.6)		76 (51.4)	72 (48.6)	

gender, age, path, and latesTNM as independent variables. The propensity score for each patient can be calculated.

Step 2 Matching: 1:1 propensity score matching based on surgical B group can be conducted through macro program. The program can be found in file (30.1.2) at <http://www.worldscientific.com/r/8981-suppl>, or refer to SAS website <http://www2.sas.com/proceedings/sugi26/p214-26.pdf>.

Step 3 Testing the balance of covariates between two groups before and after matching: The balance of important covariates between the two surgical groups before and after matching can be tested by *t*-test for continuous variables and Chi-square test for discrete variables (Table 30.3).

One can see from the results above, the distributions of covariates between the two groups are significantly different before matching but achieve a balance after that.

Step 4 After the matching process: A survival analysis is conducted for the matched data (259 cases for both A and B groups) to compare the survival time (see Program 30.2).

Program 30.2 Survival analysis using matched data.

Line	Program
01	PROC LIFETEST METHOD=PL
02	DATAFILE="H:\Experiments 30-1 after matching.XLS";
03	PLOTS=(S);
04	TIME OSmonths*statusOS(0)
05	STRATA VATS2;
06	RUN;

Experiment 30.2 Propensity score matching of non-balanced sample

To study the impact of non-balanced sampling in groups *A* and *B* on survival time, the following experiment is to generate a set of simulation data of survival time, which follows an exponentially distribution, and to explore the impact of propensity score matching on survival analysis.

Step 1 Generating exponentially distributed survival time *ht*: The covariates include x_1 , x_2 and irrelevant variable x_3 , where x_1 follows a binary distribution ($p = 0.5$), x_2 and x_3 follow the standard normal distribution. x_1 and x_2 are involved in the generation of survival time *ht*, which follows an exponential distribution with a parameter of $0.5x_1 + 0.4x_2$. The censoring indicator is a variable flag (0 indicates censoring). The dataset *A* is generated by repeat the procedure 1000 times, and it is divided into two groups by the dichotomous variables x_1 . Within each group, the data are divided into four layers ($x_3 < 0, ht < 0.25$), ($x_3 < 0, ht \geq 0.25$), ($x_3 \geq 0, ht < 0.25$), and ($x_3 \geq 0, ht \geq 0.25$) corresponding to $g = 1, 2, 3, 4$, respectively.

Program 30.3 Generating 1000 individuals randomly. Line 03 generates random a variable x_1 which follows a binary distribution; lines 04 and 05 generate the independent variables x_2 and x_3 of which both follow standard normal distribution; lines 06–11 generate the values of survival time with censoring around a constant 1.5623, the censored proportion is about 50%; lines 12–15 divide 1000 individuals into four layers ($g = 1, 2, 3, 4$) according to the values of the variables x_3 and *ht*.

Step 2 Cox proportional hazards regression: To conduct Cox regression by phreg process using random variables generated above. Flag = 0 indicates censoring data (see Program 30.4).

Program 30.3 Generating exponentially distributed survival time *ht*.

Line	Program
01	data A;
02	do i=1 to 1000 ;
03	x1= rantbl(0,0.5,0.5)-1;
04	x2=normal(0);
05	x3=normal(0);
06	z=RANUNI(0);
07	t=-log(z)*exp(-(0.5*x1+0.4*x2));
08	t2= ranuni(0)*1.5623;
09	ht=min(t,t2);
10	if t > t2 then flag=0;
11	else flag=1;
12	if x3<0 and ht<0.25 then g=1;
13	ELSE IF x3<0 and ht>0.25 THEN g=2;
14	ELSE IF x3>0 and ht<0.25 THEN g=3;
15	ELSE g=4;
16	End;
17	output;
18	run;

Program 30.4 Cox regression.

Line	Program
01	proc phreg data=A;
02	model ht*flag(0)=x1 x2 x3/selection=stepwise sle=0.05 sls=0.05 RL;
03	Output survival=s;
04	run;

Result 1: Cox regression results of the original data
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter estimate	Standard error	Chi-square	Pr > Chi-square	Hazard ratio	95% Hazard ratio Confidence limits	
x1	1	0.52056	0.08518	37.3516	<0.0001	1.683	1.424	1.989
x2	1	0.45499	0.04414	106.2562	<0.0001	1.576	1.446	1.719

Program 30.5 Unbalanced sampling.

Line	Program
01	DATA s1 s2 ;
02	SET A;
03	IF x1=1 THEN OUTPUT s1;
04	ELSE OUTPUT s2;
05	RUN;
06	proc sort data=s1;
07	by g;
08	run;
09	proc surveyselect data=s1 out=r1 method=srs samprate=(0.9,0.3,0.3,0.4);
10	STRATA g;
11	run;
12	proc sort data=s2;
13	by g;
14	run;
15	proc surveyselect data=s2 out=r2 method=srs samprate=(0.5,0.3,0.3,0.9);
16	STRATA g;
17	run;
18	data B;
19	set r1 r2;
20	run;

Step 3 Unbalanced sampling of the two datasets: $x_1 = 1$, $x_1 = 0$ indicate individual in group 1 and group 2 respectively. The sampling ratio of the four layers was (0.9, 0.3, 0.3, 0.4) in group 1 and (0.5, 0.3, 0.3, 0.9) in group 2.

Program 30.5, lines 01–05 divide the dataset *A* into two subsets *s1* and *s2* in accordance with $x_1 = 0$ and $x_1 = 1$; lines 06–08 and lines 12–14 sort the datasets *s1* and *s2* according to the stratification variables *g*; lines 09–11 and lines 15–17 use the surveyselect process to conduct unbalanced sampling in accordance with the stratification variable *g*; *srs* means sampling without replacement; lines 18–20 merge the sampling data into data set *B*, which contains about 500 individuals (here we have 501 individuals).

Step 4 Cox proportional hazards regression of the samples obtained by non-balanced sampling: For data obtained from unbalanced sampling, Cox regression is conducted by *phreg* process where *flag* = 0 indicates censoring (refer to step 2 for programs).

Result 2: Cox regression results of unbalanced sampling.
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter estimate	Standard error	Chi-square	Pr > Chi-square	Hazard ratio	95% Hazard ratio Confidence limits	
x1	1	0.90344	0.12455	52.6130	<0.0001	2.468	1.933	3.151
x2	1	0.40428	0.06121	43.6225	<0.0001	1.498	1.329	1.689
x3	1	-0.13622	0.05890	5.3481	0.0207	0.873	0.778	0.979

Step 5 Matching by propensity score: Matching the two groups in dataset *B* obtained by non- balanced sampling through propensity score matching method used above, there will be about 150 pairs of individuals (here we have 155 pairs).

Step 6 Cox proportional hazards regression of the non-balanced sample after matching of propensity score: For data obtained after propensity score matching, Cox regression is conducted through phreg process where flag = 0 indicates censoring.

Result 3: Cox regression results after propensity score matching.
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter estimate	Standard error	Chi-square	Pr > Chi-square	Hazard ratio	95% Hazard ratio Confidence limits	
x1	1	0.81626	0.15454	27.8989	<.0001	2.262	1.671	3.062
x2	1	0.38722	0.07669	25.4919	<.0001	1.473	1.267	1.712

It can be learnt from this example that the generation of survival time (*ht*) only involved covariates *x* 1 and *x* 2 (see results 1), but after unbalanced sampling of *ht* which is correlated with *x* 3. The regression by Cox model showed that all the coefficients of variables *x* 1, *x* 2 and *x* 3 were statistically significant (see results 2). Here variable *x* 3 was false positive. However, after using matching of propensity score, Cox regression showed that only coefficient of variables *x* 1 and *x* 2 are statistically significant. Hence matching of propensity score can eliminate the impact of confounder *x* 3 to a certain extent. We suggest the reader to think about the reason of the above phenomenon.

(2nd edn. Jing Gu, Qian Zhao, Jiqian Fang)

Chapter 31

Statistical Methods in Scale Development

With the change from biomedical model to biopsychosocial medical model, the clinical professionals gradually pay more attention to the patients' psychological characteristics and feelings, which are usually measured by scales. How to develop a qualified scale? Does a scale measure the expected contents? Are the results measured by the scale reliable? How to analyze the data measured by the scale? This chapter will discuss these issues.

31.1 Development of Scales

In the widest sense, scale development is a complex process involving a whole set of methods and including a framework ranging from establishment of concepts, operational definitions, development of item pool, selection of items, construction of scale, and test of psychometric properties. While in the narrow sense, it is just a process of definition of concepts and items development.

31.1.1 *Main steps of scale development*

The main steps of scale development are described by the following example on developing a Quality of Life Scale.

1. Objectives and subjects

We should firstly determine the subjects and make sure we would develop a generic scale or a specific scale for a particular population, such as the elderly people or the cancer patients. Whether the scale aims

at discrimination or evaluation is another issue need to be considered beforehand as well.

2. Working group

A nominal group and several focus groups should be organized. The members of focus groups who are responsible for raising items should be somewhat extensively selected, including physicians, nurses, carers and patients in addition to the experts.

3. Definition and decomposition

This step would be completed by focus groups. The main task of this step is to clarify the definition of the concept to be measured and the structure of it. For example, the definition of quality of life and the meaning of its subscales, domains and facets.

4. Item pool development

In this step, the items associated with the content clarified in step 3 are put forward according to their professional knowledge and personal experience independently. And the nominal group goes through a process of rewording, classification and analysis to perform an item pool.

5. The format of the response scale

There are two different formats for recording the response. One is the so-called linear analog painting technique, which prepares a line segment with 0 and 10 attached to the two ends respectively as a coordinate, and asks the subject to score his (or her) response by setting a point on it. Another is the orderly level off method, which prepares several adjectives or adverbs with equidistance describing different levels of the response and asks the subject to select one and only one level as a score to match his (or her) response.

6. Item analysis and selection

In this step, items are analyzed and screened based on the scores with statistical methods to develop a preliminary scale.

7. Pilot study

This step aims at evaluating reliability, validity, response feature, and other psychometric features of the scale through pilot study.

8. Revision

After revision on the basis of the above steps, we get the final scale.

31.1.2 *Item analysis and item selection*

1. Item analysis

Item analysis is a necessary part of pilot study. It is a process of evaluating the items from various aspects and providing evidences for item selection, including evaluation of difficulty, response feature, discrimination, representativeness, and independence.

Analysis of difficulty can be evaluated by the response rate. If an item is only answered by a few subjects, then we can conclude that it is unsuitable or difficult.

Analysis of response feature is aimed to determine the validity of options and how the subjects response to them. It is unsuitable if the responses from the subjects concentrate on some special options rather than others.

2. Item selection

Item selection is a key task in development of a scale. The importance, sensitivity, independence, representativeness, and certainty of items should be taken into consideration. The feasibility and acceptability of them should be addressed as well.

The following are the approaches of item selection from various perspectives and serve different purposes. The sample sizes are better to be larger than 100.

- (1) Subjective evaluation. This approach is to select items according to their importance to health. Doctors or patients would be asked to score the importance of each item independently (hundred-mark system would be used, while ranking is another choice when the number of the items is small). Then the items will be selected according to their mean or median scores, and those low-scored items could be abandoned. In addition, evaluations of importance should be conducted among doctors and patients separately because they might have different views.
- (2) Dispersion trend method. This method is to select items according to their sensitivity. The smaller the dispersion tendency is, the worse

ability of discrimination the item has. So the items with large discrete tendency could be selected. The indices reflecting dispersion tendency depend on the distributions and features of the scores of each item. Generally speaking, standard deviation and coefficient of variation are widely used.

- (3) Correlation coefficient method. This method is to select items according to their representativeness and independence. Through calculations and statistical tests of correlation coefficients, those having many or few related items will be selected. The former can provide more information because of their representativeness while the latter cannot be replaced by others because of their independence. The Pearson product-moment correlation coefficient, Spearman or Kendall rank correlation coefficient are widely used.
- (4) Factor Analysis. This method is to identify the factor structure of the scale firstly. Then the items with large loading and which are consistent with the factor structure could be selected. For example, if a scale for quality of life is conceived to have five main factors including physical function, psychological status, social relationship, symptoms and toxicity, environment, we can choose the items related to these five factors.
- (5) Cluster analysis. This method is also to select items according to their representativeness. Using a clustering method (hierarchical approach is commonly used), the items could be classified into several clusters. In every cluster, the item(s) mostly correlated with others on average could be selected as the representative ones.
- (6) Regression analysis. Subjects are asked to make a total score for their overall evaluation (such as quality of life). This total score could be used as a dependent variable Y . Then a multiple regression analysis between Y and items (X_1, X_2, \dots, X_p) will be conducted to select the items which have big influences on Y . One can get different numbers of important items for further choosing if different testing levels (α) are selected. This method can also be used at a domain level to select items for the specific domain.
- (7) Discriminant analysis. A good scale should have the ability of distinguish different populations (such as patients and healthy people). Given datasets from two different populations in pilot study, a

discriminant analysis could be conducted to select items which are good in distinguishing the two populations.

Any of the above-mentioned method has its own advantages and disadvantages in selecting items. The subjective evaluation method is based on subjective judgements of evaluators, and pilot studies are not necessary, while the other methods, based on measured results, need some pilot studies. The subjective evaluation method based on importance assessment can be used firstly to select among large amount of items, then put the resulted items into pilot study and select them by other methods.

Generally speaking, the correlation coefficient method, factor analysis method and clustering analysis method emphasize on relational structure of data, and select the items on the perspective of representativeness. Factor analysis method and clustering analysis method can be conducted due to the existing of correlation among items. The correlation analysis method is flexible, but do not take into account the relationship among items and the structure of scale. The results issued by different clustering methods are various and difficult to discern which is better. One can try several clustering methods at the same time, and take an overall consideration of the results. The discrete tendency method, regression analysis method and discriminant analysis emphasize the variation of structure of data, which select items on the perspective of sensitivity and importance. So the items with large variation possibly tend to be selected. To sum up, different methods are distinguished and related at the same time. These methods can be combined to select items. The items being selected initially should go through further testing of other features like feasibility, reliability, validity and so on to continually determine if being retained or abandoned.

31.2 Adopting Scale with Foreign Language

There are no more than two ways regarding to scales, which include reformulating a new one and using the developed ones. How to translate an already developed foreign scale and adjust it to the new culture background? Does the process of transformation include only translation? Will the translated foreign scales be appropriate for domestic measurement? The answer is definitely no.

31.2.1 *Translation and back-translation of scale*

1. Translation

Usually the “forward” translation is suggested to be accomplished by at least two translators who are familiar with the source language and its cultural background as well as the target language with great proficiency.

2. Back-translation

Back-translation is a key procedure to check the equivalence. It is suggested to be worked out by those differing from the forward translators and without reading the initial version of the source language. By comparing the back-translated scale with the initial scale, one may find the flaws in the translation or the points need cultural adaptation.

31.2.2 *Cultural adaptation of scale*

The process of cultural adaptation is to assess the equivalence between the new scale and the initial one. There are at least six aspects being considered in terms of equivalence.

1. Conceptual equivalence

The conceptual equivalence means whether there exist identical definition and understanding in different culture background. For instance, in the research of cross-culture health-related quality of life (HRQOL) measurement, it is basically assumed that there exists a universally acknowledged QOL definition which can be measured by an identical set of domains. Besides, it is required to maintain the same response scale cross different culture background, so as to guarantee the consistency of measurement results.

Evaluation method: the most frequently used way is to review relevant literatures from different countries or regions and grasp their definition and understanding about the concept, and then conclude whether these thoughts are equivalent. Also it is useful to consult expert or to organize focus group discussion to evaluate the equivalence. The factor analysis could be used to assess conceptual equivalence in terms of potential factor structure, loading of each item on the relevant factors.

Taking QOL scales as an example, there may be four kinds of evaluation results: the first one is that the definition of QOL and the importance degree

of each domain are the same between the two scales; the second is that the definition of QOL is the same but the importance degrees of each domain are different; the third is that the definition of QOL and the importance degree of each domain are partially the same; the last is that the definition of QOL and the importance degree of each domain are totally different.

The first kind of result shows good conceptual equivalence. The second shows that there exists conceptual equivalence and the weight coefficients could be used to reflect the different importance among domains. The third and the last demonstrate that there is no conceptual equivalence. For the third one, the different domains could be regarded as a part of cultural specificity when evaluating QOL.

2. Item equivalence

Similarly, the item equivalence between the two scales concerns whether the roles of the item are the same within the domains. For instance, the item on sleeping pill usage may generate cultural distinction, because in certain countries people do not use these pills. Another example is that using one's ability in looking after his or her garden might not be proper to measure the health situation and activity ability, because most people do not own private gardens in some countries. When the item equivalence is evaluated, one should not only consider the efficiency but also the acceptance of the item. Some items are considered impolite or offensive in certain culture background. For example, it is not proper to ask one's sex life under some circumstances. When and only when the item plays the same role in the domain with the same efficiency and acceptance, the item equivalence could be concluded.

Evaluation method: The most frequently used methods include reviewing local literature, Delphi evaluation method and focus group discussion. By reviewing relevant literature about every region, especially on anthropology and sociology, one can know more about the characteristics about local culture.

Delphi evaluation method helps to know the experts' evaluation about the item.

Furthermore, ranking the items according to their importance through focus group might make the researchers know more about the efficiency of the item. From the viewpoint of statistics, item equivalence means that an

item measures the same potential factor, also means that the correlations among items in different cultures are the same. The item response theory can be used to evaluate the property of a given item; the Cronbach's alpha reliability coefficient can be used to evaluate the inner consistency of the items.

There may be four kinds of results: the first is good item equivalence, or the items can be translated from the source language to the target language for direct use; the second is item equivalence with slight adjustment; the third one is no equivalence, and the item should be replaced by others; the fourth one is not only no equivalence but also taboos so that the item should be deleted.

3. Semantic equivalence

The semantic equivalence concerns the semantic delivery by different languages is equivalent and resulting in the same response. In general, the meaning of a word might include denotative and connotative ones. Denotation refers to that expressed by the word itself, and can be searched in the dictionary; connotation refers to the implied meaning of the specific context, which is gained by sociology and anthropology research. To achieve semantic equivalence, one must catch precise understanding about the key words in the scale before translation.

Evaluation method: the semantic equivalence can be evaluated by strictly examining the process of translation-back translation. In case that a few items can hardly be made up a semantic equivalence for some reason, one needs to replace or even delete them at the end.

4. Operational equivalence

Operational equivalence means that similar format, administration mode, time specification and measurement methodology are used.

Evaluation method: These can be checked by the focus groups and experts consultation.

For certain practical reason, some operational procedures might need to be adjusted. For example, change the administration mode of self-report by report with help of investigators or change the phone-based survey by mail-based one, etc. Under such circumstance, evidence is needed to show that the results from different administration mode are acceptable.

5. Measurement equivalence

Measurement equivalence is to guarantee the similarity of psychometric characteristics of different linguistic versions, especially the reliability, validity and responsibility.

Evaluation method: the Cronbach's α coefficient and test-retest correlation could be used to evaluate the reliability; the hypothesis testing and confirmatory factor analysis (CFA) could be used to evaluate the discriminant validity, convergent validity and construct validity; and the Item Response Theory (IRT) could be used to evaluate measurement equivalence.

There might be two equivalent levels in practice: The first level, there is a similar factor structure between different linguistic versions; the second level, relevant factor loadings are similar.

6. Functional equivalence

Functional equivalence can be defined as the overall equivalence achieved by the scales used in two or more than two kinds of cultures, which is the comprehensive outcome of all kinds of equivalence mentioned above.

There are three levels on functional equivalence: the first, each of the above-mentioned equivalence plays well and the outcomes of scale measurement are comparable and merged between different cultures; the second, there exists conceptual equivalence while other kinds of equivalence are not so ideal, then the measurement results should be transformed firstly before being compared or merged; the third, there exists no conceptual equivalence even though other kinds of equivalence are acceptable, then the results are not comparable and their implication in different cultures are distinct.

The six kinds of equivalences mentioned above aim to help us have a better understanding of the scale equivalence. Conceptual equivalence which requires researchers to consider its efficiency and necessity deliberately before introducing an existing scale is significantly important. If it is not proper to introduce an existing scale, researchers should consider establishing a new scale based on specific culture. Currently some researchers rely too much on the translation-back-translation procedure but ignore the equivalence evaluation, thus a real cross-cultural study will be difficult to perform.

31.3 The Concept and Evaluation of Validity and Reliability

31.3.1 Validity

Validity is concerned with whether a variable measures what it is supposed to measure. For instance, does an IQ test measure intelligence? Does a quality of life scale measure people's quality of life? Does a depression questionnaire measure the degree of patient's depression? These are the questions on validity, but they can never be answered with absolute certainty. Although we can never prove the validity, we can develop some indices to evaluate it. Traditionally, statisticians have distinguished four types of validity: content validity, criterion validity, construct validity, and convergent-discriminant validity. Content validity is largely a "conceptual test", whereas the other three types are empirically rooted. If a measure truly corresponds to a concept, we would expect that all four types of validity would be satisfied.

1. Content validity

Content validity is a qualitative type of validity where the domain of a concept is made clear and the analyst judges whether the measures fully represent the domain. To the extent that they do, content validity is met. An expert evaluation method can be used to evaluate the content validity.

Just as a non-representative sample of people can lead to mistaken inferences to the population, a non-representative sample of measures can distort our understanding of a concept.

The major limitation of content validity stems from its dependence on the theoretical definition. For most concepts in the social sciences, no consensus exists on theoretical definitions. In this situation the burden falls on researchers not only to provide a theoretical definition accepted by their peers but also to select indicators that fully cover its domain and dimensions. Definition of "quality of life" is just a case that lacks of consensus, leading to big differences between many existing quality of life questionnaires.

2. Criteria validity

Criterion validity is the degree of correspondence between a measure and a criterion variable, usually measured by their correlation. To assess criterion

validity, we need an objective reliable standard measure with which to compare our measure. Suppose that in a survey we ask each employee in a corporation to report his or her salary. If we had access to the actual salary records, we could assess the validity of the survey measure by correlating the reported ones and the records. In this case the employee records represent an ideal, or nearly ideal, criterion of comparison.

One may adopt the absolute value of the correlation between a measure and a criterion to assess criteria validity. Does this correlation coefficient reveal the validity of a measure? Only when the criterion measures the concept we concerned perfectly, can correlation coefficient reveal the true validity. However, the limitation of criterion validity is that for many measures no perfect criterion is available.

3. Construct validity

The evaluation of construct validity is usually completed by a factor analysis approach. To design a questionnaire, the researchers often begin with a set of theoretical relations as the bases of a concept to be measured. Then based on the observed data, a factor analysis is used to examine whether the questionnaire reflects the postulated theoretical construct, and confirm whether the researcher's hypotheses are consistent with the real data.

The main function of factor analysis is to draw some common factors from a series of variables measured by a scale. Different from the observed variables (also called manifest variables, like the score of each item in the scale), these factors are called latent variables. The latent variables are unobservable, while the relationship between them and the manifest variables can be investigated. Generally, the manifest variables are divided into several groups; the variables of each group share a common factor; the common factors reflect the structure of the whole scale. Therefore, the factor analysis not only can assess the construct validity, but also can investigate the structure of the whole scale through grouping all of the observed variables. The term of factor analysis as a whole includes exploratory factor analysis and confirmatory factor analysis. Confirmatory factor analysis is preferable than exploratory factor analysis, whenever one wants to confirm the construct validity. Relevant materials are referred to the chapter of factor analysis in this book.

4. Discriminant validity

Discriminant validity means that a well-designed scale should be able to distinguish the characteristics between certain target populations (such as “healthy subjects” and “patients”). For instance, it could be assessed through the following procedure: investigating some target populations respectively, then calculating the scores of domains and total score, and finally, a t test or analysis of variance is applied to test whether there are significant differences between the scores of the populations. If the hypothesis test shows significant differences, the discriminant validity is met.

31.3.2 Reliability

Reliability is the consistency of measurement. It is not the same as validity since we can have consistent but invalid measures. To illustrate reliability, suppose that I want to measure your level of education. I narrowly define education as completed years of formal schooling. I operationalize it by asking: “How many years of formal school have you had?” Next, I record your answer. If I had the ability to erase your memory of the question and the response you gave, I could repeat the same question and again, record your answer. Repeating this process an infinite number of times, I could determine the consistency of your response to the same question. The reliability of this education measure is the consistency in your response over the infinite trials. The greater the fluctuation across your answers, the lower the reliability of the measure is.

It is possible to have a very reliable measure that is not valid. For example, repeatedly weighing yourself on a bathroom scale may provide a reliable measure of your weight but the scale is not valid if it always gives a weight that is 5 kg too light. A more extreme example would be to obtain a measure of intelligence by asking individuals their shoes size. This may provide a very reliable measure, but it lacks validity as an intelligence measure. Thus the distinction between reliability and validity is very important.

Much of the social science literature on reliability originates in classical measurement theory from psychology. A fundamental equation of the theory is

$$x_i = \tau_i + e_i,$$

where x_i is the i th observed variable (or "test" score), e_i is the error term and τ_i is the true score that underlies x_i . It is assumed that $\text{cov}(\tau_i, e_i) = 0$ and $E(e_i) = 0$. According to classical test theory, the errors of measurement for different items are uncorrelated. The correlation between two measures results from the association of their true scores.

Reliability is defined as a ratio between the variance of true scores $\text{Var}(\tau_i)$ and the variance of observed variables $\text{Var}(x_i)$, which equals to the square of correlation coefficient between observed scores and true scores.

A number of methods have been proposed for estimating the reliability of measures. Here we review the three most common ones: test-retest, split-halves, and Cronbach's α .

1. Test-retest method

The test-retest method is based on administering the same measure for the same variable at two points in time, of which the difference should not be too long so that the subject's condition does not change during the period. A correlation analysis or hypothesis testing between scores of two tests is adopted to evaluate the reliability of the scale. When a statistical significant correlation coefficient results in the correlation analysis or none significant difference results in the hypothesis testing, the reliability is met. This method is particularly suitable for a factual scale. A correlation coefficient obtained from correlation analysis is named as test-retest reliability, of which a recommend standard is not less than 0.7.

Test-retest reliability assessment is difficult to operate in practice. First, it assumes perfect stability of the true condition. In many cases the true condition may change over time so that the difference between two tests is not simply caused by random errors. Second, memory effect is often present that the response of the first interview can influence the response in a second interview, and the latter response tends to be the same with the former one. Consequently, both too short and too long lengths of the time interval are not proposed. Many researchers make a recommendation of two to four weeks.

2. Split-halves method

When a test-retest method is impossible to operate, an alternative means is to divide all items into halves, of which the correlation coefficient r is used to calculate a reliability coefficient as the assessment of reliability.

The question is how to divide into halves. Generally, factual items are difficult to divide, for different characteristics are incomparable. Therefore, such method does not suit the kind of factual scales.

For the kind of attitude scales, the items are generally a variety of positive or negative statement around a certain theme, and the subjects are asked to make choices among statements. For example, to choose one of “very dissatisfied”, “dissatisfied”, “neither satisfied nor dissatisfied”, “satisfied” and “very satisfied”, and score them from 1 to 5. To divide all items into half, it can be based on the sequential order or parity of the item number as far as the two halves one similar in content, format and amount of items. The correlation coefficient r between scores of the halves is merely the reliability of the half scale though the reliability of the whole scale can be gauged by the Spearman-Brown Prophecy formula

$$R = \frac{2r}{1+r}. \quad (31.1)$$

A recommend standard of it is not less than 0.7.

The split-halves test is more desirable than the test-retest that it only needs the measurements at one time point and without the trouble of memory effects so that it is often cheaper and easier in performance.

The disadvantage of split-halves method is the way that the halves are allocated is somewhat arbitrary. There are many possible ways of dividing a set of items into half, and each split could lead to a different reliability estimate.

3. Chronbach's α coefficient

Split-halves reliability coefficient is established in the assumption that the variance of the scores of the two halves are equal, which is not always satisfied. If the variances are not equal, the reliability will be underestimated.

L.J. Chronbach proposed to use the coefficient α to assess reliability:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k s_i^2}{s_T^2} \right), \quad (31.2)$$

where k is the total number of items, s_i^2 is the variance of the score of item i , s_T^2 is the variance of the total score. The Chronbach's α is the most popular reliability coefficient in social science research at present. Generally,

a standard of no less than 0.7 is recommended, while some researchers preferred 0.9 or above.

When calculating the coefficient α , additional attention should be paid when the scale consists of more than one domain. In such a case, in addition to the whole scale, it is more proper to calculate the coefficient α for each domain respectively.

What both split-halves method and Chronbach's α coefficient actually assess is the internal consistency of the scale such that the former assesses the consistency between two halves of the scale, while the latter assesses the consistency among all items. This is a kind of homogeneity. If the consistency does not exist, the integration of the scores becomes unreasonable. Thus, to improve the reliability, we should pay attention to the homogeneity of statements originally when design a scale: whether the items describe a certain characteristic in the same direction, and some items that are likely to cause heterogeneity need to be ruled out.

Even though a scale has been demonstrated reliable and valid, it still cannot be marked as an effective tool if it cannot detect some subtle, clinical significant and time-dependent changes. Responsibility to change is considered as a validity, which is also called sensitivity. It means the measure must sensitively response to the change of observations when their internal or external environments change. We determine the observations under several different conditions, then examine the corresponding measuring result, to see whether there exist any differences.

Researchers often evaluate the responsibility to change by the following method: to investigate the objects prior treatment and post treatment respectively using the same scale and record the scores. If the treatment or intervention is effective, a significant difference will be observed between the two scores on average. In this case, we can use a paired-sample t -test to see whether the prior-post difference and correlation are statistical significant, and the responsibility to change is also assessed.

31.3.3 Case study

In this section the process of validation and reliability evaluation of WHOQOL-BREF will be introduced. It can help the readers to get a better understanding of basic concepts and learn how to assess reliability and validity of a scale.

The concept of health related quality of life (HRQOL) is revised from health defined by WHO. WHO makes a definition to health as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" in 1985. Based on this definition, researchers wish to come up with some new indices to assess the impact of disease and impairment on daily activities and behavior, perceived health measures and disability/functional status measures. Hence, HRQOL is proposed and it is of concerned by many researchers, which is comically viewed as the missing measurement in health. Although there is no consensus of definition for HRQOL currently, researchers reach an agreement in the content of it. Most of them think that HRQOL should include five domains: physical status and functional abilities, psychological status and well being, social interactions, economic and/or vocational status and factors, and religious and/or spiritual status. HRQOL is defined by WHO as "individuals' perception of their position in life in the context of the culture and value systems in which they live in relation to their goals, expectations, standards and concerns". It is a broad ranging concept incorporating in a complex way the persons' physical health, psychological state, level of independence, social relationships, personal beliefs and their relationships to salient features of the environment.

The following question is how to measure HRQOL? Specially designed measurements are no doubt needed, and the measurements are scales commonly. The development of a new QOL instrument requires a considerable amount of detailed work, demanding patience, time and resources. In short, a scale design contains several basic steps: establishment of concepts, operational definitions of every domain and field, item generation and selection, question formatting, preliminary study, psychometric evaluation, revise and field test.

The World Health Organization Quality of Life (WHOQOL) is an international scale developed by WHO to assess individual's HRQOL. Currently WHOQOL has two versions, they are WHOQOL-100 (contains 100 items) and WHOQOL-BREF (contains 26 items). The two scales were developed by 15 (9 added later) centers of different countries or districts with different cultural and economic backgrounds, and all steps are with the uniform leadership of WHO.

Table 31.1 Structure of WHOQOL-BREF.

Domain	Item
I. Physical health	3. To what extent do you feel that physical pain prevents you from doing what you need to do? 16. How satisfied are you with your sleep? 10. Do you have enough energy for everyday life? 15. How well are you able to get around? 17. How satisfied are you with your ability to perform your daily living activities? 4. How much medical treatment do you need to perform your daily life? 18. How satisfied are you with your capacity for work?
II. Psychological	5. How much do you enjoy life? 7. How well are you able to concentrate? 19. How satisfied are you with yourself? 11. Are you able to accept your bodily appearance? 26. How often do you have negative feelings such as a blue mood, despair, anxiety and depression? 6. To what extent do you feel your life is meaningful?
III. Social relationships	20. How satisfied are you with your personal relationships? 22. How satisfied are you with the support you get from your friends? 21. How satisfied are you with your sex life?
IV. Environment	8. How safe do you feel in your daily life? 23. How satisfied are you with your condition of living place? 12. Do you have enough money to meet your needs? 24. How satisfied are you with your access to health services? 13. How available is the information that you need in your day-to-day life? 14. To what extent do you have the opportunity for leisure activity? 9. How healthy is your physical environment? 25. How satisfied are you with your transport?
Comprehensive	1. How would you rate your quality of life? 2. How satisfied are you with your health?

According to the assumption of the WHOQOL research group, the WHOQOL-BREF contains four domains, each of which has six items. Besides, the scale includes two questions to measure the total quality of life and total health condition viewed by individuals. Ultimately, the WHOQOL-BREF contains 26 items in total (Table 31.1).

A preliminary test is conducted to assess reliability and validity of the questionnaire after its first formation. In the preliminary test, the questionnaire was administrated to no less than 300 participants in each research center. 250 subjects are patients and others are healthy people, half male and half female. Then a psychometric evaluation is carried out through the following steps.

Firstly, the validity of the scale will be evaluated, which is concerned with whether the scale measures people's quality of life. In order to assess the content validity of the scale, experts can be asked to evaluate whether the items of the scale measure people's quality of life according to the definitions of the concepts and the domains included.

The process of the scale developing indicates that WHOQOL-BREF has good content validity. Since there is no criterion scale which measures people's quality of life according to the concept defined by WHO, the criterion validity cannot be assessed.

The confirmatory factor analysis (CFA) was used to assess the construct validity of the scale with three steps. Firstly, factor structure was drawn according to the theoretical framework hypothesized during the process of scale developing (Fig. 31.1); secondly, the model was constructed according to the factor structure; finally, the goodness-of-fit between the model and data was assessed using the main indices including χ^2 and goodness-of-fit index (GFI). The value of χ^2 is sensitive to the sample size and the deviation from normal distribution. Some researchers suggest that χ^2 can be regarded as the statistic measuring the goodness-of-fit instead of a test statistic that a bigger value of χ^2 indicates bad fit. The common procedure is to compare both the χ^2 values and degrees of freedom of the two models; if relative to the decrease in degrees of freedom, the decrease in χ^2 is big enough, model with more parameters is acceptable. The values of GFI range from 0 to 1 and bigger GFI indicates better goodness-of-fit. Generally, a GFI value above 0.9 is considered good construct validity.

A CFA of the data from the preliminary test of WHOQOL-BREF was performed and the value of GFI was 0.904, which indicated an adequate model fit and good construct validity of the scale. The *t*-test was applied to compare the mean scores of physical health domain, psychological health domain, social relationships domain, and environment domain between the groups of patients and healthy people. The statistical significant difference

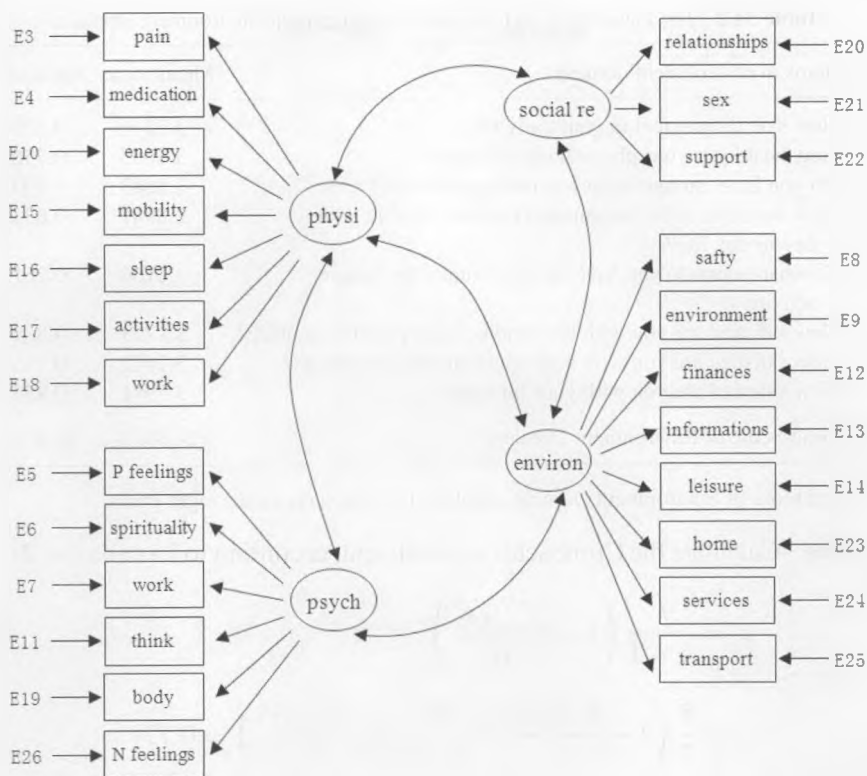


Fig. 31.1 Factor structure of WHOQOL-BREF.

($P < 0.05$) demonstrated a good discriminant validity of the scale. Above all, the validity of WHOQOL-BREF is good.

Then evaluate the reliability of the scale. The test-retest reliability was not evaluated since initially the quality of life of the objects was not measured by the scale repeatedly. The Cronbach's α coefficient was used to evaluate the reliability of the scale.

WHOQOL-BREF includes four domains: Physical Health, Psychological Health, Social Relationships, and Environment. Here, the method of calculating Cronbach's α coefficient and the split-half reliability of the Environment Domain is demonstrated as follows.

Example 31.1 The Environment Domain including eight items, the content, mean score, and variance of each item are listed in Table 31.2.

Table 31.2 The mean score and variance of each item in Environment Domain.

Items in environment domain	Mean score	Variance
1. How safe do you feel in your daily life?	3.3546	0.535
2. How healthy is your physical environment?	3.1053	0.756
3. Do you have enough money to meet your needs?	2.8643	0.707
4. How available is the information that you need in your day-to-day life?	2.8947	0.633
5. To what extent do you have the opportunity for leisure activities?	3.0166	0.761
6. How satisfied are you with the conditions of your living place?	3.1773	0.885
7. How satisfied are you with your access to health services?	3.2022	0.745
8. How satisfied are you with your transport?	3.1911	0.855
Total Score of Environment Domain*	24.8061	18.473

*Total score of Environment Domain equals to the sum score of the eight items.

Solve Calculate the Cronbach's α coefficient, according to formula (31.2),

$$\begin{aligned}\alpha &= \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k s_i^2}{s_T^2} \right) \\ &= \frac{8}{7} \left(1 - \frac{0.535 + 0.756 + \cdots + 0.855}{18.473} \right) = 0.779.\end{aligned}$$

It indicates that the Environment Domain has good reliability.

The split-half reliability can also be calculated. Eight items in Environment Domain are divided into two parts, the first part includes the first four items (items 1, 2, 3, 4) and the second part includes the last four items (items 5, 6, 7, 8). Calculate the sum of scores in the first part and denote by H_1 , and then calculate the sum of scores in the second part and denote by H_2 . The correlation coefficient between H_1 and H_2 is $r = 0.675$. Hence, the split-half reliability is

$$R = \frac{2r}{1+r} = \frac{2 \times 0.675}{1+0.675} = 0.8060.$$

The value of split-half reliability is similar to the Cronbach's α coefficient, indicating that the Environment Domain has good reliability.

Following the same method, the Cronbach's α coefficients of the Physical Health Domain, Psychological Health Domain, and Social Relationships Domain are 0.8474, 0.7919, and 0.7179 respectively. The Cronbach's

α coefficients of four domains show that the WHOQOL-BREF scale has good reliability.

31.4 Item Response Theory and Scale Evaluation

The methods of reliability and validity evaluation described above are based on the classical test theory (CTT). In addition, Item Response Theory (IRT) is commonly used in this field. IRT, formed in the early 20th century, is a new modern measurement theory which is superior to CTT. The birth of IRT changed people's research ways and presentation modes of measurement properties and changed the methods of processing scores. IRT is now widely used in education, psychology, medicine and other research fields.

31.4.1 *Concepts and models*

IRT is the general latent trait models. The relationship between examinees' item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic function (ICF) or item characteristic curve (ICC). After clarifying the characteristics of dimension (unidimensional and multidimensional), the relationship between items (independent and dependent) and pattern of responses (full or none, grade scoring, step-by-step scoring), examinees' abilities and performance indices of items can be estimated by using proper probabilistic models. The following introduction are about models for full-none performance only.

1. Model assumptions of IRT

The dependent variable is the dichotomous response (such as success/failure, refuse/accept) of the item that an examinee performs, while the independent variable is the latent trait of the examinee.

Two basic assumptions:

- (1) For items with different parameters, the shapes of ICC are different: the shape of ICC describes the relationship between the examinee's latent trait and the probability of response for the item. Different items may have different ICC shapes.
- (2) Local independence: it means that the probability of response for an item is only influenced by the examinee's ability. In other words,

after taking the examinee's ability into account, no relationship exists between the examinee's responses to different items.

Besides the two basic assumptions above, there are other assumptions for different IRT models.

2. Basic models for full-none performance

(1) Logistic Model

Logistic model rests on logistic distribution. If we use w_{is} to denote the joint action of the examinee and item parameters in the model, the probability that the examinee answers the item correctly with a given w_{is} is calculated by the equation

$$P(X_{is} = 1|w_{is}) = \frac{e^{w_{is}}}{1 + e^{w_{is}}}.$$

Table 31.2 gives 3 logistic models. For the 1-parameter logistic model, $W_{is} = \theta_s - b_i$; for the 2-parameter logistic model, $W_{is} = \alpha_i(\theta_s - b_i)$; and

Table 31.3 The 1-, 2-, and 3-parameter logistic models and their assumptions.

Model	Model equation and meaning	Assumption
1-parameter logistic model (1PL) (Rasch)	$P(X_{is} = 1 \theta_s, b_i) = \frac{e^{(\theta_s - b_i)}}{1 + e^{(\theta_s - b_i)}}$ <p>The simplest IRT model. θ_s is the latent trait of the examinee s, b_i is the item i's difficulty parameter, P is the probability that the examinee s answers the item i correctly.</p>	① Unidimensionality ② Local independence ③ Guessing is 0 ④ Same discriminant
2-parameter logistic model (2PL)	$P(X_{is} = 1 \theta_s, b_i, \alpha_i) = \frac{e^{\alpha_i(\theta_s - b_i)}}{1 + e^{\alpha_i(\theta_s - b_i)}}$ <p>Added in the item discriminant parameter. α_i is the item i's discriminant parameter, P is the probability that the examinee s answers the item i correctly.</p>	① Unidimensionality ② Local independence ③ Guessing is 0
3-parameter logistic model (3PL)	$P(X_{is} = 1 \theta_s, \beta_i, \alpha_i) = c_i + (1 - c_i) \frac{e^{\alpha_i(\theta_s - b_i)}}{1 + e^{\alpha_i(\theta_s - b_i)}}$ <p>Including both discriminant parameter and pseudo guessing parameter. c_i is the item i's pseudo guessing parameter, P is the probability that the examinee s answers the item i correctly.</p>	① Unidimensionality ② Local independence

the 3-parameter logistic model includes a pseudo-guessing parameter c_i in addition to W_{is} .

(2) Traditional Normal Ogive Models

The same as Logistic Models, the Traditional Normal Ogive Models can be divided into 2-parameter model and 3-parameter model according to the number of item parameters with the assumptions corresponding to the Logistic Models (see Table 31.4). Using the cumulative probability of normal distribution to express the probability that the examinee answers the item correctly, that is,

$$P(X_{is} = 1|Z_{is}) = \int_{-\infty}^{Z_{is}} \frac{1}{(2\pi)^{1/2}} e^{(-t^2/2)} dt,$$

where $Z_{is} = a_i(\theta_s - b_i)$.

In addition to the basic models for full-none performance, there are Polytomous IRT Models (including Unidimensional Graded Response Model, Adjusted Graded Response Model, Partial Credit Model, and Adjusted Partial Credit Model), Non-parameter Models, and Multidimensional Models.

Table 31.4 The 2- and 3-parameter traditional normal ogive models.

Model	Model equation
2-parameter model	$P(X_{is} = 1 \theta_s, b_i, a_i) = \int_{-\infty}^{a_i(\theta_s - b_i)} \frac{1}{(2\pi)^{1/2}} e^{-t^2/2} dt$
3-parameter model	$P(X_{is} = 1 \theta_s, b_i, a_i, c_i)$ $= c_i + (1 - c_i) \int_{-\infty}^{a_i(\theta_s - b_i)} \frac{1}{(2\pi)^{1/2}} e^{-t^2/2} dt$

3. Advantages of IRT (compared to CTT)

Comparing to CTT, IRT has the following advantages:

- (1) The estimation of examinee's latent trait does not depend on the specific item. IRT put the examinee's latent trait and item "difficulty" on the whole scale to perform the estimation. No matter the item is "difficult" or "easy", the estimation of the examinee's latent trait is invariant.
- (2) The estimation of the parameters of "difficulty" or "discrimination" has nothing to do with the examinee. For the same item, responses of

high-ability and low-ability examinees are fitted the same ICC, and the corresponding item parameter is unique.

- (3) The notion of information function of IRT takes the place of reliability theory, which uses the amount of information provided by item at the examinee's ability to indicate the reliability of the test. IRT avoids the assumption about "parallel tests" and gives the test precision of examinees with different abilities.

31.4.2 Application of IRT on scale evaluation

- (1) Test the construct validity of scale

Choose an appropriate model to fit the items by dimensions or by all. If the items fit well by dimensions but not fit well by all, it indicates that each dimension tests one side of the latent trait, so the scale has good construct validity.

- (2) Test the measurement bias or differential item functioning (DIF)

In the study of Quality of Life Scale, DIF is defined as: in the same condition of quality of life, the distributions of scores of the item are not the same in different groups (gender, age or country, etc.). The principle of applying IRT in DIF analysis is comparing the parameters among different groups. If the parameters are identical, it is considered as no DIF existing among different groups.

- (3) Scale development and modification

In IRT, information function is used to describe the test efficiency of a scale or an item.

Generally, for the item i , the information function is denoted as

$$I_i(\theta) = \frac{[P'_i(\theta)]^2}{P_i(\theta)Q_i(\theta)}$$

$P'_i(\theta)$ is the derivative of the item response function (the probability that the examinee answers the item correctly) with respect to θ and $Q_i(\theta) = 1 - P_i(\theta)$.

Test information function $I(\theta)$ in essence is the amount of information when using the scale to test the examinee's latent trait, which is simply the sum of the item information functions at θ . The test information function is the biggest amount of information for the whole test, no matter what kind

of scoring methods is used, which can be used in scale development and item selection. The principle of applying IRT in item selection is that the test information function achieves the desired goal by the least items.

Example 31.2 Differential item functioning analysis for WHOQOL-100 basing on the Hong Kong and Argentina data from WHOQOL Group (Yaofeng Han, Yuantao Hao, Jiqian Fang. Detection of Differential Item Functioning in Cross Cultural Analysis of the WHOQOL-100. Chinese J. of Health Statistic (2009) 26(4), 338–339).

Solution The principle and steps in applying IRT-ANOVA to DIF analysis are shown below

- (1) Merged the data from the two centers, Hong Kong and Argentina, and then estimate the parameters of each item and examinee;
- (2) Calculated the probability of the examinee n selecting k on item i (P_{nik}) by using the parameters from step (1);
- (3) Calculated the expectation and variance of the item i 's score of the examinee n by using the probability from step (2);

$$E_{ni} = \sum_{k=0}^{m_i} k P_{nik},$$

$$V[x_{ni}] = \sum_{k=0}^{m_i} k^2 P_{nik} - E_{ni}^2.$$

- (4) Calculated the standardized residual of the item i 's score of the examinee n

$$Z_{ni} = \frac{x_{ni} - E_{ni}}{\sqrt{V[x_{ni}]}} Z_{n_{cg}i}.$$

In order to carry a further ANOVA, examinees were divided into ten blocks according to the location parameter, respectively indicated by the subscript c ; and the subscript g was used to indicate the center. Therefore, the examinee could be indicated as n_{cg} , whose standardized residual could be indicated as $Z_{n_{cg}i}$

$$Z_{n_{cg}i} = \frac{x_{n_{cg}i} - E_{n_{cg}i}}{\sqrt{V[x_{n_{cg}i}]}}.$$

Table 31.5 Results of DIF analysis using IRT-ANOVA for social domain of WHO-QOL-100.

Item	Research centers			Blocks			Research center \times Blocks			DIF
	<i>F</i>	<i>df</i>	<i>P</i>	<i>F</i>	<i>df</i>	<i>P</i>	<i>F</i>	<i>df</i>	<i>P</i>	
F13.1	1.844	1	0.1747	1.724	9	0.0790	1.057	9	0.3920	
F13.2	2.248	1	0.1340	2.050	9	0.0311	0.932	9	0.4956	
F13.3	2.384	1	0.1229	3.575	9	0.0002	0.975	9	0.4584	
F13.4	134.502	1	<0.001	1.629	9	0.1020	1.070	9	0.3825	✓
F14.1	0.266	1	0.6064	0.805	9	0.6115	1.880	9	0.0511	
F14.2	5.379	1	0.0202	2.304	9	0.0144	0.371	9	0.9489	
F14.3	2.743	1	0.0979	1.542	9	0.1281	0.970	9	0.4629	
F14.4	12.218	1	0.0005	3.752	9	0.0001	2.285	9	0.0153	✓
F15.1	0.053	1	0.8185	4.048	9	<0.0001	1.494	9	0.1450	
F15.2	0.052	1	0.8197	3.724	9	0.0001	7.261	9	<0.0001	✓
F15.3	15.291	1	0.0001	3.671	9	0.0001	0.785	9	0.6300	✓
F15.4	24.424	1	<0.0001	5.246	9	<0.0001	1.511	9	0.1387	✓

Significant level $\alpha = 0.10$.

- (5) Performed ANOVA to the standardized residual. Dependent variables were centers and blocks, and then we analyzed their main effects and interaction effect of centers and blocks. Either main effects or interaction effect having statistical significance indicated the item had DIF between Hong Kong and Argentina (see Table 31.5).

31.5 Computer Experiments

Experiment 31.1 Reliability analysis of a scale Program 31.1 is used for analyzing the results of Example 31.1.

31.6 Exercises and Experiments

1. What are validity and reliability? What is the purpose of validity and reliability evaluation?

Program 31.1 Reliability analysis of the results of Example 31.1.

Line	Program	Line	Program
01	DATA cronbach;	08	5 5 4 3 4 4 4 5 34
02	INPUT q1-q8 total @@;	09	PROC MEANS mean var maxdec=4;
03	CARDS;	10	VAR q1-q7 total;
04	1 1 1 1 2 1 1 1 9	11	RUN;
05	1 4 1 1 1 4 1 1 14	12	PROC CORR data=cronbach alpha nocorr;
06	...	13	VAR q1-q8;
07	5 5 4 3 4 5 4 4 34	17	RUN;

2. What are the common methods of validity and reliability evaluation? What characteristics do they have?
3. How to evaluate the validity and reliability of a new scale? What are the steps in detail?

(2nd edn. Yuantao Hao, Nanqiao Cai)



Chapter 32

Statistical Methods for Data from Genetic Epidemiological Study

Genetic epidemiology is a relatively new discipline that seeks to elucidate the role of genetic and environmental factors in the occurrence of disease in population. The surge in the field of genetic epidemiology has been accompanied by the explosion in molecular techniques, the increasing sophistication of statistical methods and the emergence of molecular epidemiology. In this chapter, some basic concepts and theories of linkage analysis and genetic association analysis are introduced.

32.1 Basic Concepts

32.1.1 *Genetic terminology*

In the human somatic cell, there are 23 pairs of chromosomes, including 22 pairs of autosomes and one pair of sex chromosomes. Females have two of the same kind of sex chromosome (XX), while males have two distinct sex chromosomes with difference in shape and size (XY). The chromosomes determine the cell differentiation, cell function and the development of the human body, behavior and intelligence. The observable trait is called phenotype, such as height, weight, blood type and disease status. Gene, a molecular unit of heredity, is a functional part of DNA, which determines various biological traits. It is estimated that there are about 10,000 nuclear genes in human beings. Each gene has its specific location in the chromosome, called locus. A locus sometimes represents a perceptive DNA marker or a DNA fragment with polymorphism. Any of the alternative forms of a gene that may occur at a given locus are termed as alleles. Alleles are often denoted by letters or numbers, such as $A, a, B, b, 1, 2$ or 3 . The pair of



Fig. 32.1 Locus, homozygote and heterozygote.

alleles at a locus is referred to as the genotype; such as AA, Aa or aa. The individual with identical alleles is called a homozygote, such as AA or aa. The individual with different alleles is called a heterozygote, such as Aa.

In Fig. 32.1, X_1 and X_2 are a pair of homologous chromosomes. M1 and M2 are two different genes and they have specific locations in the chromosome, locus M1 and locus M2. Genes M1 and M2 may have two different alleles, that is A, a, and B, b. The individual in Fig. 32.1 has two different alleles in M1, that is, genotype Aa, so the individual is called a heterozygote at locus M1; at locus M2, the two alleles are the same, with genotype BB, so the individual is called homozygous at locus M2. The proportion of alleles at a locus in a population is called the gene frequency; for example, $P(A) = 0.3$ means that 30% of alleles at this locus in this population are A. The total of gene frequencies of all different alleles at a certain locus is 1. The genotypic frequency is the proportion of individuals carrying a certain genotype in a population; for example, $P(Aa) = 0.3$ means that 30% of individuals in the population carries the genotype Aa.

32.1.2 Hardy–Weinberg equilibrium

Random mating is defined as any female that has the same chance of mating with any male. Accordingly, the probability of mating type is the product of female genotypic frequency and male genotypic frequency; for example $P(AA \times Aa) = P(AA)P(Aa)$.

Assume a locus has two possible alleles A and a, and the frequencies of allele A and allele a are $P(A) = p_A$, $P(a) = p_a$. If three genotype frequencies for the pair of alleles in certain generation of a population are

$$P(AA) = p_A^2, \quad P(aa) = p_a^2, \quad P(Aa) = 2p_Ap_a, \quad (32.1)$$

then under random mating, there will be no change in either the allele frequencies or the genotypic frequencies in the next generation. This is an

equilibrium status in the population, and it is called Hardy–Weinberg equilibrium (HWE). In a large random mating population, without immigration, selection and mutation, the allele and genotypic frequencies will not change from one generation to the next.

If the genotype frequencies in certain generation do not satisfy condition (32.1), the next generation will be close to the equilibrium under random mating.

In genetic analysis, Pearson χ^2 test is used to test the hypothesis of HWE. The null hypothesis is that the study population is in HWE, and the test statistic is

$$\chi^2 = \sum \frac{(O - E)^2}{E}, \quad (32.2)$$

df = Number of phenotype – Number of alleles at a locus.

In Eq. (32.2), O is the observed genotypic count, E is the expected genotypic count. For the locus with two possible alleles and under the hypothesis of HWE, the expected frequencies of genotype in a random sample with n individuals are

AA	Aa	aa
np_A^2	$2np_A p_a$	np_a^2

The values of \hat{p}_A and \hat{p}_a can be estimated by:

$$\hat{p}_A = \frac{2n_{AA} + n_{Aa}}{2n}, \quad \hat{p}_a = \frac{2n_{aa} + n_{Aa}}{2n}, \quad (32.3)$$

where n_{AA} , n_{Aa} , n_{aa} are the observed numbers of the individuals with genotype AA , Aa and aa , respectively.

Example 32.1 In a genetic study of hypertension, 197 individuals were randomly selected from a population. Their genotypes of angiotensin-converting enzyme (ACE) were analyzed and the observed frequencies are

AA	Aa	aa
26	93	78

The null hypothesis of “this population is in HWE” is tested.

$$\hat{p}_A = \frac{2n_{AA} + n_{Aa}}{2n} = 0.3680, \quad \hat{p}_a = 0.6320,$$

$$\hat{E}_{AA} = 197 \times (0.3680)^2 = 26.68,$$

$$\hat{E}_{Aa} = 2 \times 197 \times 0.3680 \times 0.6320 = 91.63,$$

$$\hat{E}_{aa} = 197 \times (0.6320)^2 = 78.69,$$

$$\chi^2 = \frac{(26 - 26.68)^2}{26.68} + \frac{(93 - 91.63)^2}{91.63} + \frac{(78 - 78.69)^2}{78.69} = 0.0439,$$

$$\nu = 1.$$

Since $P = 0.8346 > \alpha = 0.05$, the null hypothesis cannot be rejected; we conclude that the population is in HWE.

32.1.3 Linkage and linkage equilibrium

From Mendel's second law, if two genetic loci are on different chromosomes, the transmission of alleles at one locus is independent of that at another locus (the recombination rate $\theta = 1/2$). However, if the two genetic loci are close together, the alleles that are paternal or maternal in the origin tend to transmit together to an offspring. This phenomenon is called as linkage. The closer the two loci are, the smaller the probability for crossing over. The recombination fraction for two linked loci is less than $1/2$.

If the alleles of two loci are randomly combined, these two loci are called linkage equilibrium. For example, there are two loci and each has two possible alleles, A, a and B, b , respectively. Their allele frequencies are

$$P(A) = p, \quad P(a) = q, \quad P(B) = u, \quad P(b) = v,$$

where $p+q = 1, u+v = 1$. When these two loci are in linkage equilibrium, the probability of joint haplotype equals the product of individual allele frequencies of two loci.

$$P(AB) = P(A)P(B) = pu, \quad P(Ab) = P(A)P(b) = pv,$$

$$P(aB) = P(a)P(B) = qu, \quad P(ab) = P(a)P(b) = qv.$$

Under random mating, there are nine two-locus joint genotype and their frequencies are

$AABB$	$AABb$	$AAbb$	$AaBB$	$AaBb$	$Aabb$	$aaBB$	$aaBb$	$aabb$
p^2u^2	$2p^2uv$	p^2v^2	$2pq u^2$	$4pq uv$	$2pq v^2$	q^2u^2	$2q^2uv$	q^2v^2

The genotypic probabilities of the next generation will be the same as the current generation.

If the alleles of two loci are not randomly combined, we call these two loci in linkage disequilibrium and δ is often used as the parameter of linkage disequilibrium

$$\delta = P(AB) - P(A)P(B), \quad (32.4)$$

$\delta = 0$ means linkage equilibrium between two loci and $\delta \neq 0$ means linkage disequilibrium, thus

$$\begin{aligned} P(AB) &= P(A)P(B) + \delta, & P(Ab) &= P(A)P(b) - \delta, \\ P(aB) &= P(a)P(B) - \delta, & P(ab) &= P(a)P(b) + \delta. \end{aligned} \quad (32.5)$$

Besides δ , δ' and r (or r^2) are also useful measures of linkage disequilibrium. δ' is defined as

$$\begin{aligned} \delta' &= \frac{\delta}{|\delta|_{\max}}, \\ |\delta|_{\max} &= \begin{cases} \min(P(A)P(b), P(a)P(B)), & \text{if } \delta > 0, \\ \min(P(A)P(B), P(a)P(b)), & \text{if } \delta < 0, \end{cases} \end{aligned} \quad (32.6)$$

δ' is sometimes called standardized δ . Obviously, $|\delta'| \leq 1$. r is defined as

$$r = \frac{P(a_1a_2) - P(a_1)P(a_2)}{\sqrt{P(A)P(a)P(B)P(b)}} = \frac{+(-)\delta}{\sqrt{P(A)P(a)P(B)P(b)}} \quad (32.7)$$

or

$$r^2 = \frac{\delta^2}{P(A)P(a)P(B)P(b)},$$

where $a_1 \in \{A, a\}$, $a_2 \in \{B, b\}$, and the sign in (32.7) is determined by a_1, a_2 , see (32.5). It is easy to show that r is actually the Pearson correlation coefficient of alleles a_1 and a_2 .

If there are two or more loci that are in linkage disequilibrium in a population, the linkage equilibrium will not be approached by random mating of one generation. It needs n generations ($n \rightarrow \infty$) to be close to linkage equilibrium or δ close to 0. The speed of approaching joint equilibrium depends on the recombination fraction θ . Let δ_0 be the initial linkage disequilibrium parameter, the linkage disequilibrium parameter after n generations is

$$\delta_n = (1 - \theta)^n \delta_0.$$

It is obvious that the linkage disequilibrium decreases quickly when the linkage is weak (θ is close to $1/2$); otherwise, more generations should be passed to approach linkage equilibrium. Therefore the value of linkage disequilibrium is the evidence of linkage in some extent.

32.1.4 Hereditary mode

Assuming a disease locus with disease allele D and a normal allele d , the prevalence of the individuals with genotype DD , Dd or dd is called penetrance of DD , Dd or dd , denoted by f_{DD} , f_{Dd} , f_{dd} respectively. The penetrance is actually a conditional prevalence, i.e. $P(\text{Affected}|\text{genotype})$. Usually, $1 \geq f_{DD} \geq f_{Dd} \geq f_{dd} \geq 0$. When the population is in HWE, the prevalence will be

$$P_A = P(\text{Affected}) = q^2 f_{DD} + 2q(1 - q)f_{Dd} + (1 - q)^2 f_{dd}, \quad (32.8)$$

where q is the frequency of allele D .

The hereditary mode is determined by the penetrances. If $1 \geq f_{DD} = f_{Dd} > f_{dd} \geq 0$, it is called dominant hereditary mode. In the special case when $f_{DD} = f_{Dd} = 1$, $f_{dd} = 0$, it is called complete dominant mode. Similarly, it is called recessive hereditary mode if $1 \geq f_{DD} > f_{Dd} = f_{dd} \geq 0$, and called complete recessive mode in the special case when $f_{DD} = 1$, $f_{Dd} = f_{dd} = 0$. It is called additive if $1 \geq f_{DD} \geq f_{Dd} \geq f_{dd} \geq 0$, $f_{Dd} = \frac{1}{2}(f_{DD} + f_{dd})$, and multiplicative if $1 \geq f_{DD} \geq f_{Dd} \geq f_{dd} \geq 0$, $f_{Dd} = \sqrt{f_{DD}f_{dd}}$.

To infer the hereditary mode, we should collect the pedigree and segregation information. The segregation analysis is commonly used to infer the hereditary mode. The inferences are based on the segregation information derived from the collected pedigree data. The details are omitted here.

32.2 Linkage Analysis

32.2.1 Introduction

The linkage analysis investigates whether or not two loci physically locate near one another on the same chromosome. The linkage analysis is one of the most important methods used to localize disease gene in human genome. The classical concept of linkage is the alleles from two linked loci (physically close) tend to segregate together, that is, they are passed from parent to child as a single unit. This phenomenon of cosegregation in a family deviates from Mendel's second law of independent assortment. The most possible biological explanation of linkage is that these two loci are physically very close in the same chromosome so that they are passed from parent to child as a single unit. Elston (1981) thought that the linkage of a known marked gene and a putative gene for a disease is considered the highest level of statistical evidence that the disease is due to a genetic mechanism. The alleles cosegregated due to linkage between two loci in one family may be different from the alleles in another family. For example, for a disease linking with ABO blood type locus, the disease allele might link with allele *A* in one family and might link with allele *B* in another family. Since the cosegregation phenomenon due to linkage is only observable within families, the family data or data from biologically related subjects are necessary for detecting linkage. Although the allelic association (linkage disequilibrium) can be detected by general population studies, the genetic linkage cannot be detected and the recombination fraction between two loci cannot be estimated by this kind of study. Allelic association is a property of alleles, while linkage is a property of loci. They are two different but related concepts. The linkage is one cause of allelic association but the allelic association is not totally caused by linkage.

The measure for linkage between two loci is the recombination fraction θ . It describes the genetic distance between two genes in a chromosome and the distance of two loci. The recombinant is that the haplotype of an individual is different from the haplotypes of his (or her) father or mother. The non-recombinant is that the haplotype of an individual is the same as that of one of his (or her) parents. As shown in Fig. 32.2, the parents' mating type is $ab/ab \times AB/ab$, the first son is a recombinant (Ab/ab) and the second son is a non-recombinant (AB/ab). The recombination is due to

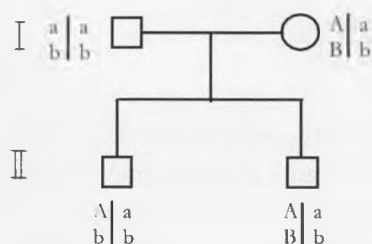


Fig. 32.2 Demonstration of recombinant and non-recombinant.

the exchange of non-sister chromatids from the homologous chromosome during the meiosis of chromosome and the genes located on some chromosome are separated from each other. The chance of recombination between two loci is proportional to the distance of two loci. The closer the two loci are, the less likely that a cross-over will occur between them. The frequency of recombination between two loci is called recombination fraction, presented as θ . If two loci are far apart and segregate independently, then $\theta = 1/2$; and if two loci are identical and are actually one locus, then $\theta = 0$. The range of the recombination fraction is $0 \leq \theta \leq 1/2$. In genetics, 1% recombination fraction is called 1 genetic distance, that is, 1 centimorgan. 1 centimorgan is about the distance of 1,000,000 base pair (1000 kb) on chromosome. There are two types of statistical methods for linkage analysis: model-based and model free.

32.2.2 The LODS method

The log-odds score (LODS) method is based on the maximum likelihood ratio test and is considered a model-based procedure. Usually, assume the mode of inheritance, the number of alleles and the penetrance of each genotype are known for the LODS method. LODS is the logarithm of a ratio between the probability of a given family when two loci are linked according to a recombination fraction θ and the probability of the family without linkage ($\theta = 1/2$), that is,

$$Z(\hat{\theta}) = \log_{10} \frac{L(\hat{\theta})}{L(\theta = 0.5)}. \quad (32.9)$$

The values of $\hat{\theta}$ lie between 0 and 0.5. Recombination fraction $\hat{\theta}$ can be estimated by maximizing Z , the LODS. Since in Eq. (32.9) the denominator

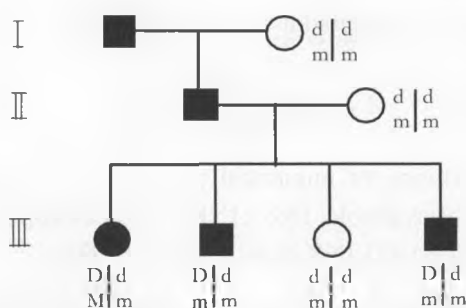


Fig. 32.3 Demonstration of a three-generation family.

is a constant and the numerator is the likelihood function for a given family, the estimated $\hat{\theta}$ value is a maximum likelihood estimation of θ .

LODS method is subject to a kind of sequential experiment method. Therefore, the LODS value for each family can be added up together according to the same recombination fractions. Generally, when $Z > 3$, it is considered as an evidence of linkage ($p < 0.0001$). By experience, $\theta \leq 0.10$ means a close linkage, $\theta \geq 0.20$ means a loose linkage, and $0.10 < \theta < 0.20$ means a median linkage.

Figure 32.3 depicts the occurrence of an autosomal dominant disease in a three-generation family. The disease is decided by the locus with two possible alleles D or d and the marker alleles are M or m . The black symbols indicate the subject is affected in the graph. The penetrances are

$$P(\text{Affected}|DD) = P(\text{Affected}|Dd) = 1, \\ P(\text{Affected}|dd) = 0.$$

Both grandmother and mother are homozygous with genotype dd at the disease locus and the genotype is mm at the marker locus. The marker genotype for father is Mm . Since the father must receive a dm from grandmother, and he is affected, the haplotype for the father must be DM/dm whatever the grandfather's genotype is DD or Dd at the disease locus. Since the mother is homozygous at both disease and marker loci, each of the four children must receive a dm from her. Therefore we may deduce the haplotype for all children as in Fig. 32.3. Only the second son belongs to recombinant based on the two locus genotype and linkage phases of parent and children. Based on the specific recombination fraction $\hat{\theta}$, the estimated Z value for

this pedigree may be calculated as:

$$Z = \log_{10} \frac{\hat{\theta}(1 - \hat{\theta})^3}{0.5(1 - 0.5)^3}.$$

Now many software or numerical procedures for LODS method of linkage analysis are available, like LINKAGE (Lathrop, 1984), MENDEL (Lange, 1988), and FASTLINK (Cottingham, 1993).

Example 32.2 In order to lock the location of disease-related gene of Machado–Joseph disease (MJD) on 14th chromosome, Wang G.X. *et al.* (1997) used 13 micro satellites DNA markers on 14th chromosome to do a linkage analysis in four MJD high risk families. There are a total of 61 members in these four families including 15 affected cases. The Mlink procedure in the software Linkage (5.22 version) was used to do two loci linkage analysis. It is assumed that the genotype frequency of MJD is 0.000002 (referring to the results of epidemiological survey in Japan). The recombination fractions for man and woman are considered as the same. Under the assumption of the autosomal chromosome dominant inheritance, the LODS values are estimated when the recombination fractions are 0, 0.01, 0.05, 0.1, 0.15, 0.2 and 0.3. The results are listed in Table 32.1.

Table 32.1 LODS of two loci linkage analysis between MJD gene and 13 micro satellite DNA markers on 14th chromosome.

Locus	Recombination fraction (θ)							Z_{\max}	θ
	0.000	0.010	0.050	0.100	0.150	0.200	0.300		
D14S59	0.73	0.72	0.66	0.60	0.52	0.43	0.24	0.73	0.00
D14S55	0.00	0.00	0.06	0.01	0.11	0.12	0.10	0.12	0.20
D14S67	−∞	1.99	2.42	2.35	2.12	1.81	1.08	2.42	0.05
D14S48	3.03	2.98	2.76	2.45	2.11	1.74	0.99	3.03	0.00
D14S291	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.00
D14S280	2.18	2.14	1.96	1.74	1.50	1.26	0.76	2.18	0.00
AFM343vfl	3.06	3.13	3.09	2.83	2.47	2.05	1.12	3.13	0.01
D14S81	4.91	4.82	4.42	3.91	3.38	2.84	1.69	4.91	0.00
D14S265	−∞	−1.98	−0.73	−0.30	−0.12	−0.03	0.02	0.02	0.29
D14S62	0.32	0.49	0.75	0.81	0.77	0.69	0.44	0.81	0.10
D14S65	−∞	−1.89	−0.46	0.07	0.30	0.38	0.31	0.38	0.22
D14S45	−∞	−2.85	−1.47	−0.90	−0.59	−0.39	−0.15	0.00	0.48
D14S51	−∞	−0.66	0.01	0.23	0.29	0.29	0.20	0.30	0.17

The LODS values (Z_{\max}) for the markers D14S48, AFM343vf1 and D14S81 are greater than 3, and the recombination fractions ($\hat{\theta}$) for them are less than 0.1. Therefore, we may consider that these three loci are closely linked to the disease-related locus of MJD.

32.2.3 The ASP method

The LODS method of linkage analysis strictly depends on the specific inheritance mode and is not proper when the inheritance mode of interested traits or disorders is not clear. In these situations, other non-parametric methods of linkage analysis are appropriate. The affected sib pair method (ASP) suggested by Penrose is introduced here.

Assume that G and T represent the main trait (disease) and measuring trait (genetic marker). Generally, when one of the parents is affected and the genotype of measuring trait is Tt , and another is not affected and the genotype of measuring trait is tt , their mating types may be supposed as: $GT/gt \times gt/gt$ or $Gt/gT \times gt/gt$. In this situation, the phenotype of children may be one of four combinations: $G-T$, $G-tt$, $ggT-$ and $ggtt$. Therefore there are ten types of sib pair in sib group (Table 32.2).

These ten types of sib pairs may be classified into four groups. In the first group, the traits of two members of sib pair are totally the same (type 1–4), no recombinant. In the second group, the first trait of two members of sib pair is the same and the second trait is different (type 5, 6), there is

Table 32.2 Ten types of sib pair combinations.

	sib1	sib2
Type 1	GT	GT
Type 2	Gt	Gt
Type 3	gT	gT
Type 4	gt	gt
Type 5	GT	Gt
Type 6	gT	gt
Type 7	GT	gT
Type 8	Gt	gt
Type 9	GT	gt
Type 10	Gt	gT

recombinant. In the third group, the first trait of two members of sib pair is different and the second trait is the same (type 7, 8), there is recombinant. In the fourth group, the traits of two members of sib pair are all different (type 9, 10), no recombinant.

If locus of G is linked to that of T , whatever the linkage phases of double heterozygote is matching ($GT/gt \times gt/gt$) or not ($Gt/gT \times gt/gt$) the frequencies of the first and fourth groups are larger than those of the other two. Table 32.2 may be regrouped as Table 32.3.

In Table 32.3, n_1, n_2, n_3 and n_4 are the frequencies of sib pairs in groups 1, 2, 3 and 4 respectively. A chi-square test for 2×2 table is used to decide if there is a linkage between G and T loci. The null hypothesis is H_0 : there is no linkage between G and T

$$\chi^2 = \frac{N(n_1n_4 - n_2n_3)^2}{(n_1 + n_2)(n_3 + n_4)(n_1 + n_3)(n_2 + n_4)}, \quad \nu = 1. \quad (32.10)$$

When the chi-square value is large, the linkage between G and T is suggested.

Example 32.3 In order to decide if the locus of antibody A links to the locus of MN blood type, and antibody A and antibody M of sib from 10 families are tested. The results are listed in Table 32.4.

Table 32.3 2×2 table of four groups of sib pairs.

Group of sib pair	GT (or gt)	Gt (or gT)	Total
GT (or gt)	n_1	n_2	$n_1 + n_2$
gT (or Gt)	n_3	n_4	$n_3 + n_4$
Total	$n_1 + n_3$	$n_2 + n_4$	N

Table 32.4 Antibody A and M of sib pairs from ten families.

Antibody	Number of families where sibs come from									
	1	2	3	4	5	6	7	8	9	10
A	++++	++	---	+--+	+--	+++	--	---	+--	++
M	++++	++	--+	+--+	---	+++	--	++-	+--	++

Table 32.5 The phenotypes of sib pairs.

Types of sib pairs	$A + M +$ (or $A - M -$)	$A + M -$ (or $A - M +$)	Total
$A + M +$ (or $A - M -$)	16	5	21
$A - M +$ (or $A + M -$)	4	3	7
Total	20	8	28

The sibs in each family may be matched to sib pair and in total 28 pairs are matched. As in the first family, four sibs may be matched to form $4(4-1)/2 = 6$ sib pairs. The phenotypes of 28 pairs are listed in Table 32.5.

When it is tested for H_0 : there is no linkage between A and M ,

$$\chi^2 = \frac{28(16 \times 3 - 5 \times 4)^2}{21 \times 7 \times 20 \times 8} = 0.9333$$

$P = 0.3340$, according to $\alpha = 0.05$, H_0 is not rejected. The results suggest that there is no linkage between antibody A and M .

32.3 Genetic Association Analysis

Genetic association studies aim at detecting the association between one or more genetic polymorphisms and a trait. Once an allele of a gene is over represented in a case population relative to the control, it may be established that such an allele of the gene is associated with the studied disease. Population-based case-control design and family-based designs such as the case-parent triad designs are often used for genetic association studies.

32.3.1 Population base association analysis

The association can be demonstrated, if it exists, by comparing allele frequencies at the marker locus in random samples of unrelated patients and controls. Therefore, unlike the linkage studies requiring data from the whole families of ASPs, genetic association studies can use population-based case-control design. In this simple case, familiar methods such as χ^2 tests of association, logistic regression, and odds ratios may be suitable.

Pearson χ^2 test statistic,

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad (32.11)$$

where O is the “observed frequency”, and E is the “expected frequency”. There are two kinds of χ^2 tests, genotype-based χ^2 test and allele-based χ^2 test. The tests are executed using a contingency table analysis with the rows representing the binary disease status and the columns representing the g observed genotype classes or k observed allele classes respectively. The genotype-based χ^2 test compares the genotype frequencies, and the allele-based χ^2 test compares allele frequencies at the marker locus in random samples of unrelated cases and controls. It is assumed that there is an additive or multiplicative allele effect on the disease susceptibility in the allele-based χ^2 test. Both tests assume HWE in the combined sample of cases and controls. When H_0 , no association between the marker and affection status of a particular disease, holds, the genotype-based and allele-based χ^2 statistics have an asymptotic χ^2 distribution with degrees of freedom (df) $g - 1$ and $k - 1$ respectively. For the biallelic markers, the genotype-based and allele-based χ^2 statistics are asymptotically χ^2 distributed with $df = 2$ and $df = 1$ respectively.

Example 32.4 In a genetic association study of diabetes, two random samples with unrelated 366 diabetes patients and 390 normal subjects were recruited respectively. The genotyping results for a candidate single nucleotide polymorphism (SNP) marker are shown in Table 32.6.

The genotype-based χ^2 test statistic

$$\chi^2 = \sum \frac{(O - E)^2}{E} = 11.268,$$

$df = 2$, $P = 0.004$. Therefore there is significant difference in genotype distribution between case and control groups, and an association between diabetes and the SNP marker is suggested. Furthermore, we can calculate

Table 32.6 Genotype frequencies for Case and control groups.

Group	Genotype			Total
	11	12	22	
Case	107	198	61	366
Control	159	181	50	390

Table 32.7 Allele frequencies for the case and control groups.

Group	Allele		Total
	1	2	
Case	412* (56.28%)	320 (43.72%)	732 (100%)
Control	499 (63.97%)	281 (36.03%)	780 (100%)

*412 = $2 \times 107 + 198$, similar for other cells.

the odds ratio (*OR*) for genotype 12 and 22, with 11 as reference level, resulting in 1.626 and 1.813 respectively.

The allele-based contingency table is shown in Table 32.7. In a similar way, we obtain $\chi^2 = 9.325$ with $df = 1$ and $P = 0.002$, and the *OR* for allele 2 relative to 1 is 1.379. The results suggest that there is a significant difference between the case and control groups for the allele frequencies.

The general genotype-based χ^2 test is performed regardless of the underlying hereditary mode. If the genotype risks are additive, the genotype-based χ^2 test will not be as powerful as allele-based χ^2 test which is tailored to this scenario, because of the smaller sample size and larger degree of freedom. If allele 1 is assumed a risk allele for a biallelic marker, one way to improve the power to detect the dominant (or recessive) risks is to count the heterozygotes with genotype 12 into homozygotes with genotype 11 (or 22 respectively) since the heterozygotes have the same risk with homozygote 11 (or 22 respectively). And then a 2×2 table rather than 2×3 table is constructed and a Pearson 1-*df* test can be applied.

For complex traits, it is widely thought that the heterozygote risk is constrained to lie within the range defined by the two homozygote risks, that is $f_{AA} \geq f_{Aa} \geq f_{aa}$, where *A* is a risk allele. The widely used dose-response model, the Cochran–Armitage trend test (also known as the Armitage test) is also used in testing a marker for association with a disease locus. The table for Armitage test is set up as Table 32.8.

The statistic for Cochran–Armitage trend test is given as

$$Z = \frac{\sum_{i=0}^2 t_i (Sr_i - Rs_i)}{\sqrt{\frac{RS}{N} \left[\sum_{i=0}^2 t_i^2 n_i (N - n_i) - 2 \sum_{i=0}^1 \sum_{j=i}^2 t_i t_j n_i n_j \right]}} \stackrel{H_0}{\sim} N(0, 1) \quad (32.12)$$

Table 32.8 Genotype distribution for case and control groups.

Group	Risk allele count			Total
	0	1	2	
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

or

$$\chi_{CA}^2 = Z^2 \stackrel{H_0}{\sim} \chi_1^2, \quad (32.13)$$

where t_0, t_1, t_2 are pre-specified weights for three genotypes according to the hereditary mode. For example, the additive allele effect model describes the linearity of the genotype-phenotype relationship for the binary trait, and t_0, t_1, t_2 can be set as 0, 1, and 2. Similarly, $(t_0, t_1, t_2) = (0, 1, 1)$ and $(0, 0, 1)$ for dominant and recessive model respectively. The linear trend test statistic corresponding to additive model with $(t_0, t_1, t_2) = (0, 1, 2)$ is

$$Z = \frac{\sqrt{N}[N(r_1 + 2r_2) - R(n_1 + 2n_2)]}{\sqrt{RS[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]}}. \quad (32.14)$$

The linear trend test for the data from Table 32.7 shows $Z = 3.128$ and $P = 0.0018$.

In the presence of the effects of covariates, the stratified χ^2 test or logistic regression can be used to detect and control their effects. The stratified χ^2 test can be used when there exist a few categorical covariates only; otherwise, the logistic regression is a better choice.

32.3.2 Family-based association analysis

The association between a disease and a genetic marker can arise from confounding by underlying stratification and admixture (substructure) within the population. The population stratification can occur in case-control or other population-based designs. It is important to make comparisons between cases and controls within homogeneous subpopulations as far as possible. The family-based designs have been proposed to counteract confounding due to population stratification. The best-known family-based

Table 32.9 Combinations of transmitted and nontransmitted marker alleles A and a among $2n$ parents of n affected children.

Transmitted allele	Nontransmitted allele		
	A	a	Total
A	n_{11}	n_{12}	n_{1+}
a	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	$2n$

design is the case-parent triad design. The case-parent triad design and transmission disequilibrium tests (TDTs) proposed by Spielman (1993) suggested to collect case-parent trios. The alleles or genotypes transmitted to affected individuals are compared with untransmitted alleles or genotypes, providing that a control sample is inherently matched to the case sample with regard to population structure. TDT also has the advantage that it does not require data either on multiple affected family members or on unaffected sibs.

Suppose that we have a sample of n case-parent trios. In these families there will be a total of $4n$ parental marker alleles, $2n$ of which are transmitted and $2n$ of which are not transmitted. The data on marker alleles in the affected children can be set up as in Table 32.9.

n_{11} , n_{22} in Table 32.9 are the numbers of the homozygous parents. The transmitted allele from homozygous parent is the same as the nontransmitted allele. n_{12} (or n_{21}) is the number of heterozygous Aa parents who transmit A (or a) and do not transmit a (or A) to the affected offspring. TDT suggests that the association exists if a significant difference is observed between n_{12} and n_{21} . TDT proposed by Spielman is also referred to as McNemar's χ^2 test.

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \quad (32.15)$$

Note that only the data from heterozygous parents should be used in the test since the homozygous parents provide no information about the transmission disequilibrium.

Example 32.5 The data used here were from Spielman *et al.* (1993) assembled from 94 families with two or more insulin-dependent diabetes mellitus (IDDM) children. The candidate marker is the region of tandem-repeat DNA (5' flanking polymorphism [5'FP]) adjacent to the insulin gene on chromosome 11p. The heterozygous parents transmitted 124 alleles (78 class 1 alleles and 46 class X alleles) to their diabetic offspring.

Here n_{11} , n_{22} are 78 and 46 respectively when TDT is used.

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} = \frac{(78 - 46)^2}{124} = 8.26,$$

with degree of freedom $df = 1$ and $P = 0.004$. This finding suggests that the 5'FP contributing to IDDM susceptibility.

TDT has been extended to a marker locus with more than two alleles (Spielman, 1996; Sham, 1995). For the TDT in the general case when the marker has $m (\geq 2)$ alleles, the details on the transmitted/nontransmitted and shown in Table 32.10.

n_{ij} in Table 32.10 is the number of parents with genotype ij who transmit i and do not transmit j to the affected offspring. Sham (1995) pointed out that the test of symmetry which tests the symmetry of the TDT table can be used to detect the transmission disequilibrium. The statistic

$$T_s = \sum_{i < j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}, \tag{32.16}$$

Table 32.10 Combinations of transmitted and nontransmitted marker alleles 1, 2, ..., m among $2n$ parents of n affected children.

Transmitted allele	Nontransmitted allele				Total
	1	2	...	m	
1	n_{11}	n_{12}	...	n_{1m}	n_{1+}
2	n_{21}	n_{22}	...	n_{2m}	n_{2+}
...
m	n_{m1}	n_{m2}	...	n_{mm}	n_{m+}
Total	n_{+1}	n_{+2}	...	n_{+m}	$2N$

Table 32.11 TDT table for 154 parents from 77 trios.

Transmitted allele	Nontransmitted allele			Total
	1	2	3	
1	9	17	22	48
2	18	20	13	51
3	19	23	13	55
Total	46	60	48	154

is asymptotically χ^2 distributed with $df = m(m - 1)/2$, under H_0 . In the special case when $m = 2$, T_s is the TDT statistic proposed by Spielman.

Example 32.6 In an association study in German population between neural tube defect (NTD) and gene Mthfr (Stegmann *et al.*, 1999), 77 affected-parent trios were collected, see Table 32.11.

$$\begin{aligned}
 T_s &= \sum_{i < j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} \\
 &= \frac{(18 - 17)^2}{18 + 17} + \frac{(22 - 19)^2}{22 + 19} + \frac{(23 - 13)^2}{23 + 13} \\
 &= 3.026, \quad df = 3
 \end{aligned}$$

and $P = 0.388$. Therefore no association for the marker Mthfr is concluded.

32.4 Computerized Experiments

Experiment 32.1 Test of H-W equilibrium The data used in the test for HWE analysis is from Example 32.1. In Program 32.1, lines 01–13 are used to build the dataset HW with 26, 93 and 78 observations with genotypes AA, Aa, and aa, respectively. HW test for the built dataset via allele procedure is performed in lines 14–16.

Experiment 32.2 Genetic association analysis with case-control design Three statistical methods, the genotype case-control test, Armitage's trend test and allele case-control test are used for testing the association between

Program 32.1 Test for Hardy-Weinberg equilibrium.

Line	Program	Line	Program
01	DATA HW;	09	END;
02	DO id=1 TO 26;	10	DO id=120 TO 197;
03	allele1="A"; allele2="A";	11	allele1="a"; allele2="a";
04	OUTPUT;	12	OUTPUT;
05	END;	13	END;
06	DO id=27 TO 119;	14	PROC ALLELE data=HW;
07	allele1="A"; allele2="a";	15	VAR allele1 allele2;
08	OUTPUT;	16	RUN;

Program 32.2 Genetic association analysis with case-control design.

Line	Program	Line	Program
01	DATA geno;	11	DATA allele;
02	INPUT r c f @@ ;	12	INPUT r c f @@ ;
03	CARDS;	13	CARDS;
04	1 1 107 1 2 198 1 3 61	14	1 1 412 1 2 320
05	2 1 159 2 2 181 2 3 50	15	2 1 499 2 2 281
06	;	16	;
07	PROC FREQ data=geno;	17	PROC FREQ data=allele;
08	WEIGHT f;	18	WEIGHT f;
09	TABLES r*c/NOCOL	19	TABLES r*c/NOCOL
	TREND CHISQ;		NOPERCENT CHISQ;
10	RUN;	20	RUN;

a marker and a disease locus. The data used in the analysis are from Example 32.4. In Program 32.2, lines 01–06 are used to construct the 2×3 contingency table with 2 rows representing binary disease status and 3 columns representing the 3 genotype classes. The FREQ procedure in lines 07–10 is used to perform classical χ^2 test (genotype-based χ^2) test and Cochran–Armitage linear trend test with the option TREND. 2×2 allele based contingency table is built by lines 11–16. The rest lines are for the allele-based χ^2 test.

Experiment 32.3 TDT analysis The data used in the analysis is from Example 32.6. In Program 32.3, lines 01–07 are used to construct the 3×3 transmitted/nontransmitted table shown in Table 32.12. The FREQ

Program 32.3 TDT analysis.

Line	Program	Line	Program
01	DATA A;	08	PROC FREQ;
02	INPUT r c f @@ ;	09	WEIGHT f;
03	CARDS;	10	TABLES r*c/NOROW NOCOL
04	1 1 9 1 2 17 1 3 22	11	CHISQ AGREE;
05	2 1 18 2 2 20 2 3 13	12	RUN;
06	3 1 19 3 2 23 3 3 13	13	PROC PRINT;
07	;		RUN;

Table 32.12 Genotype fistribution of 155 affected-parent trios.

Genotype of parents	Genotype of the affected child		
	CC	CT	TT
CC × CC	19		
CC × CT	4	4	
CC × TT		7	
CT × CC	36	30	
CT × CT	15	15	6
CT × TT		10	7
TT × TT			2

procedure with option "AGREE" in lines 8–10 is used to detect the symmetry of the TDT table.

32.5 Practice and Experiments

1. For a random mating population, estimate the genotype proportion in the next generation: (0.25,0.10,0.65); (0.30,0.0.70); (0.60,0.40).
2. Test whether the genotype frequencies in the following population are HW or not and calculate the equilibrium proportion for those non-HW population: (50%, 0%, 50%); (36%, 15%, 49%); (9%, 10%, 81%); (45%, 45%, 10%).
3. What are the HWE and linkage equilibrium?
4. For the biallelic system, show that the LD coefficient r is just the Pearson correlation coefficient.

5. How to choose the LODS method and ASP method in a linkage analysis? What is the relationship between recombination fraction and the genetic distance? What is the range of recombination fraction value?
6. The genotype information of 155 affected-parent trios collected for an family-based association study are shown Table 32.12. Construct a transmitted/nontransmitted table and perform a TDT analysis.

(1st edn. Qing Liu, Zongli Xu, Caixia Li, Jiqian Fang; 2nd edn. Caixia Li, Jiqian Fang)

Chapter 33

Statistical Methods in Bioinformatics

Bioinformatics was a new interdisciplinary subject which dated from late 1980s, due to the rapid growth genome sequencing data. Our ultimate purpose was to disclose innovation from hug biological data and to find out what the living creatures were going through by the strategy for analyzing and processing with data which challenge adventurous biologists and mathematicians.

Bioinformatics works upon enormous database which is comprised of two-level database. The original data from experiment or those with a few simply treat only, e.g. data arrange and annotate, are stored into the first level database; and the second level database is the developed database of the first, which is not only derived, but also theoretically analyzed for certain aims based on the first level database. The Genbank, EMBL and DDBJ, etc. are noted as the first level nucleic acid databases, as well as some protein sequence databases, e.g. SWISS-PORT and PIR, etc. and some protein structure databases, such as PDB. Many second level databases have been created by different features for several objects, of which GDB is a human genome database, TRANSFAC is a transcription factor binding sites database, and SCOP family is a protein structure classification database, etc.

To create and implement a new analytical tool is one of the core issues of bioinformatics, so as to get more biological information from the above-mentioned databases. Hence, Statistics figures one important in Bioinformatics, for instance, rip and rigorous multi-sequence comparison methods, testing methods for large-scale and multi-level complex statistical analysis methods, etc. Those include certain amounts of statistics skills. This chapter will give an introduction of it, and will also describe statistics applied in bioinformatics in various aspects.

33.1 Sequence Alignment Methods

In molecular biology research, normally, it has to look up from databases to find out similar homologous sequences for a new detective nucleotide sequence or its translated amino acid sequence, in order to speculate the possibly belonged family and function of unknown gene sequence. To an amino acid sequence, a ready known three-dimensional structure homologous protein could be found potentially to speculate the spatial structure of unknown sequence. Then, database is essential to bioinformatics as an important tool, either database search or database query.

Protein similarity search usually draws on heuristic algorithm, with some optimization criteria, can calculate the “almost” correct answer immediately. The BLAST, FASTA and Smith–Waterman based dynamic programming algorithm are the three popular algorithms for database searching.

33.1.1 *BLAST database search tool*

BLAST, an abbreviation for Basic Local Alignment Search Tool, is an algorithm for comparing primary biological sequence information (Altschul *et al.*, 1990;1997), of which the main idea is that there are often high-scoring segment pairs (HSP) contained in a statistically significant alignment. BLAST searches for high scoring sequence alignments between the query sequence and sequences in the database using a heuristic approach that approximates the Smith–Waterman algorithm. Then the search starts from the segment to both ends to detect a well-matched segment as long as possible.

33.1.2 *Sequence similarity*

Sequence similarity is a very direct quantitative relation, a lot of measurements to describe, but in common definition are distance and similarity. Distance, the definition is allocate a weight value to each potentially mutated in biological evolution, and defined two sequences such that one of them can be transformed into another by series variation. Therefore, the distance between two sequences is calculated as the minimal sum value of those variation’s weight values. Similarity is also based on two defined sequences, scoring (weight valuing) the segment pairs while matched at

each point. The similarity is calculated as the maximal sum value of those scores (weight values).

Assigning a sequence to apply database for similar retrieval, the following situations may result:

- (1) The sequence completely matches with a sequence in database (exactly the same).
- (2) The sequence is similar to those interesting sequence (oncogenes or growth factors, cytokines, etc.) clearly.
- (3) The sequence is similar to the general featured sequence in database (Cytochrome C, rib nuclease).
- (4) Very faint similarities between two sequences, e.g. two sequences' residue resemble at 15–25%. Doolittle called this situation between similar and dissimilar twilight zone (common occurred situation).
- (5) Not match at all.

Researcher must ensure whether the incorrect occurred in sequencing process and searchable database is up-to-date before confirming that a new protein has been found through sequence comparison, which must proceed with extreme caution. If it is a new sequence with no similarity to any other, then it is a unique sequence which can be used as probes for withdrawing genes containing such DNA sequence from genomic library. This kind of unique sequence is barely able to find, as the nucleic acid and protein sequence databases keep expanding. Any long enough sequence may find some similarity sequences partly by matching in sequences database.

In most cases, whether the high scored sequences in waiting list are truly related to the one detected during database searching require further tests to pair detected sequence with those in waiting list to full-scale alignment comparisons and statistical tests.

Despite the sequence registered residues' likelihood can reflect the similarity size, there still exist influence factors, e.g. length of compared sequence, the number of brought in empty positions, etc. Therefore, it does not come along with similarity size as direct proportion. The result has to pass through statistical tests after sequence comparison to identify whether it has statistical significance.

33.1.3 Sequence similar statistical tests

The Monte Carlo simulation method is a very direct and simple way to identify whether the score of a pair of the sequence full-scale alignment was statistically significant. It randomly alters the symbols of sequences and then calculates a new score of full-scale alignment by the same procedure and variable parameters. Such a process is repeated approximately 100 times, resulting in 100 mean values and standard deviations. Assuming random sequence full-scale alignment score satisfies a standard normal distribution, then decision making could be based on the Z -value. If the Z -value equal to 3SD, 4SD or 5SD units then the random occurrence of full-scale alignment scores with its probability are 10^{-3} , 10^{-5} or 10^{-7} . In the circumstance that Z -value is over 5SD, two compared proteins full-scale alignment scores could be recognized such that their theoretical mean difference is not zero and these two proteins were homologous evolution. For Z -values between 3SD to 5SD, and other evidence to prove similarities (functional similar) between two proteins, they could be identified as homologous evolution, but not for Z -values less than 3SD.

Many sequence comparison software with Z -value calculating program are able to evaluate the full-scale alignment level directly, e.g. PIR protein analysis, ALIGN and RELATE program computing the results and Z -values of sequence alignment. Therefore, human $\alpha - 2$ micro globulin and cattle/goat immunoglobulin sequence registering Z -value is calculated as 5.83SD. The IDEAS and SEQDP software have their own programs to calculate Z -value, SEQDP and RDF2 respectively.

The Karlin-Altschul formula for BLAST scoring test: Monte Carlo stimulation method assumes a large number of random registering as normal distributed. However, every scored randomized variable is the maximum (optimal registering) among large number of score data. Hence normality assumption is not very rational, it can be detected and seen clearly by fitting the scored data into a normality curve. Vingron and Watterman has popularized and applied the formula to local sequence alignment score statistical test formula in which the sequence length is one of the parameters. And Karlin and Altschul had analyzed the distribution from BLAST. For two sequences a and b , BLAST found out high scoring pairs (HSP) were $a_i \cdots a_{i-k}$ and $b_j \cdots b_{j-k}$. And the matching regions score is defined as the sum of similar values $S(x, y)$ coming

from PAM250 matrix which is $\sum_{i=0}^k S(a_{i+l}, b_{j+l})$. To calculate and report HSPs' separation points, Karlin and Altschul generated an optimal HSP test formula. The probability of optimal HSP scoring $H(a, b)$ surpasses the threshold t is:

$$P(H(a, b) > t) \approx 1 - e^{-rnmp^t}. \quad (33.1)$$

Note that r and p can be computed directly or from an equation, and m and n are the length of two sequences. It indicates that the HSPs score is approximately Poisson distributed. Hypothesize that each of the symbol's expectation score is negative. Therefore, positive scored HSP are rare events, the occurring of randomized variable HSP score over threshold is nearly distributed as Poisson. (It is based on the mean value λ .)

$$\sum_{i=0}^{\infty} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda}. \quad (33.2)$$

In this way, the maximum score below threshold t with the probability is $p(\lambda, 0) = e^{-\lambda}$, and the probability of score surpasses the threshold is $(1 - e^{-\lambda})$. In the Karlin-Altschul formula, $rnmp^t$ is equivalent to λ , the expectation of the number with score above HSP threshold t is $rnmp^t$.

33.2 The Data Acquisition and Standardization of Gene Expression Patterns

Gene chip is used to measure the genic expression levels or determine relative abundance of nucleic acid sequences in the target by probetarget hybridization which is usually detected and quantified by fluorophore-labeled targets. Every gene on a standardized chip can obtain an expression ratio, that under a certain condition, the ratio of expression level of such a gene compared to it expressed in control. Usually, \log_2 (ratio) is used to describe the amplitude of gene expression up- or down-regulated.

Gene expression pattern analysis has been a major subject in bioinformatics study, as well as a technically difficult subject. After transforming into mathematical problems, the analysis mission is to find out statically significant frame including global pattern and local pattern from the expression matrix.

Table 33.1 The differentially expressed genes between neural tube cell tumor and malignant glioma sample group.

No.	Genes	Brain_medulloblastomas (MD)			Brain_malignantgliomas (MG)		
		Sample 1	...	Sample 10	Sample 1	...	Sample 10
Gene1	RPS23	14.6621	...	14.7091	13.2274	...	13.3266
Gene2	SFRS3	12.9033	...	12.0275	10.9643	...	11.3182
Gene3	ZIC1	9.4644	...	13.6175	9.9696	...	11.0983
Gene4	RPL39	14.2960	...	14.6199	11.7454	...	12.7507
Gene5	KIAA0182	11.4974	...	11.6760	9.9470	...	10.9078
...
Gene111	RAB31	9.8799	...	10.0444	11.0274	...	11.4131
Gene112	LOC64204	12.3817	...	10.6534	11.9193	...	12.1503

Example 33.1 Study the biological functions of CNS of embryonal carcinoma in brain tissue, using gene chip to get microarray DNA expression data, and the dataset comes from CNS experiment network station. (<http://www.broad.mit.edu/mpr/CNS>) Each chip has located 7129 genes and 42 tumor tissue samples which consist of 10 medulloblastomas, 10 malignant gliomas, 10 AT/RTs, 8 PNETS and 4 normal cerebella samples.

112 statistically significant expression genes have been screening out from medulloblastomas and malignant gliomas in experiments. See the specific expression from Table 33.1 (① The genes are in the row and the samples are in the column, such samples may be different tissues, environments or events, etc. ② The gene is a sample of cluster analysis, the sample is the variable of cluster analysis.)

The format of gene expression data is given to facilitate further analysis.

Solution output results

Obs	Name	x1	x2	x3	...	x18	x19	x20
1	RPS23	14.66	14.02	14.80	...	14.14	13.22	13.32
2	SFRS3	12.90	11.65	11.97	...	11.27	11.14	11.32
...
111	RAB31	9.88	10.14	10.28	...	11.60	12.07	11.41
112	LOC64204	12.38	11.81	10.85	...	13.34	13.00	12.15

Data acquisition and standardization of gene expression patterns: each chip can get two wavelength channels scanning images to two fluorescent dyes after competitive hybridization experiment. Many chip scanners are fixed with graphical analysis software, which can calculate fluorescence intensity from both sample point and its background in the range of two-channel after pixels gridded, and then transform the image information into computable digital information.

It required standardizing relative fluorescence intensity for each channel after image processing. Different markers, detection efficiency to distinct fluorescent marks, or original concentration of sample RNA caused systematic error will correct while standardizing. The existing standardization methods mainly include reference point standardization, the total strength standardization, local weighted linear regression standardization and local mean standardization, etc. Each method has its own characteristics, due to difference of density expressional gene chips and practical quality of chip experiments to select suitable methods. As there are many reasons to cause systematic errors, it makes chip standardization more complicated. As errors can only be reduced by single method, but not completely eliminated, it is impossible to have chip standardization.

33.3 Differentially Expressed Genes Screening

The differentially expressed genes have statistical difference expression levels among several experimental groups called 'significant gene (or statistically significant genes)'. Usually, the expression level which is doubled or halved (i.e., $\text{Log}_2(\text{ratio}) \geq 1$ or ≤ -1) can be a standard to detect different expressions. Theoretically, differentially expressed genes can be screened by only one trial. Due to experimental errors, repeated experiments are required to inspect and verify.

Controlling multiple testing error rates (false positive rate) and guaranteeing high screening efficiency are essential for differentially expressed genes screening. Researchers had proposed a variety of methods for solving the problem of microarray data for differentially expressed genes screening, including SAM (Significance Analysis of Microarrays) is applied for differentially expressed genes screening regardless of research design and data type with gene expression profiles. And more methods, such as two samples *t*-test, Bonferroni correction, BH etc. can be used.

The Bonferroni, Sidak and Hochberg corrections are able to keep FWER and FDR at a very low level, while the number of screened differentially expressed genes is relatively small. However, they are inapplicable for data analysis of gene expression profile screening differentially expressed genes. Grouped *t*-test in the same sample size and variance can filter out the largest number of differentially expressed genes, but it may ineffectively control FWER and FDR levels and selected too many false positives differentially expressed genes. The simulation test found that both SAM and BH in screening differentially expressed genes number, false positive number, and, FWER and FDR are very nearly the same, both have selected plenty of differentially expressed genes and have controlled multiple testing error rates.

Example 33.2 In order to explore the different degree of varicocele patients' Notch1 testicular tissue in gene expression, 38 cases of varicocele patients were treated including varicocele I degree 10 cases, II degree 12 cases and III degree 16 cases. The Notch1 gene expression level has been determined from testicular tissue (Table 33.2). Can be the differentially expressed genes screened out via Bonferroni correction grouped *t*-test?

Table 33.2 Notch1 in gene expression level from 38 cases of different degrees of varicocele patients' testicular tissue.

I degree (<i>n</i> = 10)	II degree (<i>n</i> = 12)	III degree (<i>n</i> = 16)
0.552	0.109	0.058
0.513	0.101	0.046
0.451	0.116	0.056
0.451	0.125	0.052
0.559	0.127	0.066
0.619	0.122	0.042
0.502	0.112	0.042
0.527	0.106	0.049
0.563	0.104	0.045
0.505	0.127	0.057
	0.122	0.031
	0.123	0.069
		0.051
		0.038
		0.059
		0.043

Solution The example explored Notch1 gene expression differences during three degree conditions from varicocele patients' testicular tissue (Table 33.3).

The result of Notch1 gene expression level comparison under different condition degree: $F = 977.42$, $P < 0.0001$, and it suggested that the gene present different expressions in degrees (see Table 33.5).

It suggested that the Notch1 gene expression differed significantly in III degree condition varicocele patients' testicular tissue.

Table 33.3 The expression level of the three groups.

Patient's condition	Case	Mean	Standard deviation
I degree	10	0.5242	0.0518
II degree	12	0.1162	0.0094
III degree	16	0.0503	0.0102

Table 33.4 ANOVA table for three groups.

Sources	df	SS	MS	F	P
Model	2	1.4935	0.7468	977.42	<0.0001
Error	35	0.0267	0.0008		
Total	37	1.5202			

Table 33.5 The differentially expressed genes screened by Bonferroni correction grouped *t*-test.

(I) Patient's condition	(J) Patient's condition	Mean Difference (I-J)	Std. Error	Sig.	95% CI	
					Lower	Upper
I degree	II degree	0.408033*	0.011835	0.000	0.37827	0.43779
	III degree	0.473950*	0.011142	0.000	0.44593	0.50197
II degree	I degree	-0.408033*	0.011835	0.000	-0.43779	-0.37827
	III degree	0.065917*	0.010555	0.000	0.03937	0.09246
III degree	I degree	-0.473950*	0.011142	0.000	-0.50197	-0.44593
	II degree	-0.065917**	0.010555	0.000	-0.09246	-0.03937

33.4 Cluster Analysis of Gene Expression

Cluster analysis, the most popular method of gene expression analysis, is aimed at classifying genes from perspective on functional expression. From a mathematical standpoint, the obtained gene groups by cluster analysis are of similar mathematically attribute among internal group members, but different from other group members. From a biological viewpoint, clustering analysis implying biological meaning or basic hypothesis is that the gene expression spectra are similar within internal group and they may be functionally the same. However, the functionally same encoded genes (such as phosphorylation of other proteins) from products may not always share similar transcriptional pattern.

In contrast, the functionally different genes may have similar expression profiles by coincidence or because of random disturbance. Despite the occurrence of many unexpected circumstances, the numbers of functionally related genes have very similar expression profiles under a group of correlated conditions, in particular, gene co-regulated by a common transcription, or products constituted protein complex are the same, or regulated in the same pathway. Thus, in practice, similar gene expression profile can be clustered to infer the function of unknown genes.

Cluster analysis is a generally used method in pattern recognition and data mining, which is an effective knowledge-based method. It is used extensively in gene expression data analysis and it mainly includes hierarchical clustering, K-means and self-organizing map networks, etc.

33.4.1 Hierarchical clustering

System clustering, also known as hierarchical clustering, is a simple method and the results are easily visualized. It has become one of the most widely used methods in gene expression data analysis, such as yeast and human gene expressions. However, many hierarchical clustering methods remain a potential problem that the strict phyletic evolutionary tree reflects gene expression pattern improperly, which contains multiple special paths.

Example 33.3 Use hierarchical cluster to analyze 112 differentially expressed genes (data source from Example 33.1).

Solution The purpose of this study is to classify the differentially expressed genes and to study its biological function upon the classification.

To classify 20 samples, as many as 112 differentially expressed genes from two tissues, the WARD hierarchical clustering method is applied here to meet the requirements (other methods include Euclidean distance method and focus of the average clustering method, etc.). Definition of inter-cluster distance by hierarchical clustering hinged on the distances between samples (i.e., genes). Firstly, merge together two closest cases into a category among n cases (individual cases) and recalculate the inter-cluster distance. Then decide which case is to be merged with which other case (or category has been merged). This process keeps repeating till all cases are merged into a large category. Finally, the result is plotted as a clustering tree directly reflecting the clustering process.

(1) Output:

Part I

The default statistics including mean, standard deviation, skewness, kurtosis and coefficient peaks, as shown in Table 33.6.

Table 33.6 Ward minimum variance cluster analysis.

Variable	Mean	Std Dev	Skewness	Kurtosis	Bimodality
x1	11.6789	1.6784	0.0691	-1.5206	0.6432
x2	11.4111	1.3716	0.00112	-1.2289	0.5394
x3	11.9545	1.3776	0.1575	-0.9693	0.4849
x4	11.8535	1.6379	-0.0731	-1.2992	0.5636
x5	11.4632	1.4720	0.2337	-1.2157	0.5649
x6	11.6921	1.5913	-0.00822	-1.2684	0.5512
x7	11.7336	1.5742	0.0443	-1.3489	0.5779
x8	11.3426	1.3862	0.4209	-0.6403	0.4819
x9	11.6874	1.7173	0.0713	-1.1983	0.5333
x10	11.7971	1.5526	0.1005	-1.3061	0.5685
x11	11.4179	1.3259	0.1901	-0.9611	0.4884
x12	11.0575	1.1949	0.6088	-0.0260	0.4484
x13	11.2553	1.1733	0.2184	-0.4442	0.3971
x14	11.2844	1.1630	0.0730	-0.7639	0.4335
x15	11.4966	1.2682	0.2198	-0.5213	0.4092
x16	11.4721	1.2056	0.00731	-0.6646	0.4135
x17	11.6705	1.4102	0.00486	-0.7738	0.4331
x18	11.5110	1.5016	-0.0489	-1.1947	0.5309
x19	11.2287	1.2546	0.5630	-0.0384	0.4326
x20	11.2720	1.2410	0.2570	-0.8539	0.4783

Part II

The eigenvalues of the covariance matrix (Table 33.7), difference values between two vertical adjacent eigenvalues, variance ratio, and the cumulative variance ratio.

Table 33.7 The eigenvalues of the covariance matrix.

	Eigenvalue	Difference	Proportion	Cumulative
1	25.6004378	16.8781129	0.6389	0.6389
2	8.7223249	7.2474690	0.2177	0.8565
3	1.4748559	0.5969479	0.0368	0.8933
4	0.8779080	0.2190825	0.0219	0.9152
5	0.6588255	0.2434937	0.0164	0.9317
6	0.4153318	0.0236515	0.0104	0.9421
7	0.3916803	0.0750759	0.0098	0.9518
8	0.3166043	0.0348680	0.0079	0.9597
9	0.2817363	0.0795238	0.0070	0.9668
10	0.2022124	0.0313723	0.0050	0.9718
11	0.1708401	0.0117839	0.0043	0.9761
12	0.1590562	0.0179848	0.0040	0.9800
13	0.1410715	0.0083331	0.0035	0.9836
14	0.1327384	0.0172594	0.0033	0.9869
15	0.1154790	0.0068077	0.0029	0.9898
16	0.1086713	0.0102010	0.0027	0.9925
17	0.0984703	0.0098862	0.0025	0.9949
18	0.0885841	0.0218081	0.0022	0.9971
19	0.0667760	0.0186661	0.0017	0.9988
20	0.0481100		0.0012	1.0000

Root-Mean-Square Total-Sample Standard Deviation = 1.415481

Root-Mean-Square Distance between Observations = 8.952286

Part III

Cluster processing as shown in Table 33.8.

Table 33.8 Cluster processing.

NCL	Clusters joined		FREQ	SPRSQ	RSQ
111	RPS6KA1	RPS14	2	0.0001	1.00
110	FBLN1	CLUL1	2	0.0002	1.00
109	RPS8	RPL11	2	0.0002	1.00

Table 33.8 (Continued)

NCL	Clusters joined		FREQ	SPRSQ	RSQ
108	RPL18	RPS5	2	0.0002	0.999
107	HLA-DRA	CD74	2	0.0002	0.999
106	NPM1	RPL13A	2	0.0002	0.999
105	RPS3A	RPS18	2	0.0002	0.999
104	RPL19	RPL24	2	0.0002	0.998
103	RPL35A	RPS15A	2	0.0003	0.998
102	RPS29	RPS11	2	0.0003	0.998
101	NFIB	HNRPA1	2	0.0003	0.998
100	RPL9	RPL10A	2	0.0003	0.997
99	RPS9	RPS28	2	0.0003	0.997
98	RPL18A	RPLP2	2	0.0003	0.997
97	ALCAM	SOX2	2	0.0003	0.996
96	CL108	RPS21	3	0.0003	0.996
95	RELN	NNAT	2	0.0003	0.996
94	CL96	RPS19	4	0.0003	0.995
93	PTOV1	HMGB2	2	0.0003	0.995
92	GNB2L1	CL102	3	0.0003	0.995
91	RPL14	SLC25A6	2	0.0004	0.994
90	RPL27	CL99	3	0.0004	0.994
89	RPL34	RPS17	2	0.0004	0.994
88	RPL21	RPL17	2	0.0004	0.993
87	RPL39	RPL32	2	0.0004	0.993
86	SPARCL1	PEA15	2	0.0005	0.992
85	CL100	CL109	4	0.0005	0.992
84	CL106	CL98	4	0.0005	0.991
83	KIAA0182	CCNG1	2	0.0005	0.991
82	SFRS3	PTMA	2	0.0005	0.990
81	ID2B	ATP1B2	2	0.0005	0.990
80	CL105	RPS27A	3	0.0005	0.989
79	CL97	OLIG2	3	0.0005	0.989
78	CL91	CL94	6	0.0006	0.988
77	SYT11	RAB31	2	0.0006	0.988
76	SNRPD2	RPS7	2	0.0006	0.987
75	CL90	CL88	5	0.0006	0.986
74	SLC1A3	PON2	2	0.0006	0.986
73	CL101	CL110	4	0.0007	0.985
72	CL75	CL103	7	0.0007	0.984
71	KIAA0367	CL81	3	0.0007	0.984
70	CL92	CL104	5	0.0007	0.983
69	CL87	CL80	5	0.0007	0.982

Table 33.8 (Continued)

NCL	Clusters joined		FREQ	SPRSQ	RSQ
68	CL78	RPS16	7	0.0008	0.981
67	CL82	H2AFZ	3	0.0008	0.981
66	CL84	CL70	9	0.0008	0.980
65	CL83	CL93	4	0.0008	0.979
64	ZNF238	CL95	3	0.0009	0.978
63	TMSL8	NPTX2	2	0.0009	0.977
62	FHL1	DDR1	2	0.0009	0.976
61	CL74	CL107	4	0.0009	0.975
60	PCDHGC3	CL86	3	0.0010	0.974
59	CL71	ID4	4	0.0010	0.973
58	CL67	HMG2N1	4	0.0010	0.972
57	CL77	FEZ1	3	0.0010	0.971
56	CL89	CL85	6	0.0010	0.970
55	APOE	CRYAB	2	0.0010	0.969
54	AQP4	C3	2	0.0011	0.968
53	TUBB3	SOX4	2	0.0011	0.967
52	PLP1	NTRK2	2	0.0012	0.966
51	CL58	TMSB10	5	0.0012	0.965
50	MAB21L1	NEUROD1	2	0.0012	0.963
49	CL69	H3F3A	6	0.0013	0.962
48	CL61	MT1M	5	0.0014	0.961
47	INSM1	STMN2	2	0.0014	0.959
46	MYCN	PAX6	2	0.0014	0.958
45	CST3	LOC64204	2	0.0016	0.956
44	SPARC	HLA-A	2	0.0016	0.955
43	CL57	TCF12	4	0.0016	0.953
42	CL63	RBP1	3	0.0017	0.951
41	CL76	MGP	3	0.0017	0.950
40	CL59	CL79	7	0.0017	0.948
39	PMP22	CL48	6	0.0018	0.946
38	PTPRZ1	GPM6B	2	0.0018	0.944
37	CL111	CL73	6	0.0019	0.943
36	RPS23	CL49	7	0.0019	0.941
35	CL65	NPTX1	5	0.0019	0.939
34	CL43	HTRA1	5	0.0020	0.937
33	CL60	CL62	5	0.0021	0.935
32	CL56	CL66	15	0.0021	0.933
31	CL53	CD24	3	0.0022	0.930
30	CL54	SPP1	3	0.0022	0.928
29	CL68	CL72	14	0.0022	0.926

Table 33.8 (Continued)

NCL	Clusters joined		FREQ	SPRSQ	RSQ
28	CL47	CL50	4	0.0025	0.923
27	CL51	C5orf13	6	0.0027	0.921
26	CL55	MT2A	3	0.0031	0.918
25	CL28	CL64	7	0.0032	0.914
24	CL33	CL45	7	0.0033	0.911
23	CL35	CL42	8	0.0033	0.908
22	CL30	SRPX	4	0.0033	0.905
21	CL40	CL52	9	0.0038	0.901
20	CL26	GFAP	4	0.0038	0.897
19	CL34	SCG2	6	0.0040	0.893
18	CL27	CL41	9	0.0043	0.889
17	CL23	CL46	10	0.0051	0.884
16	ZIC1	CL31	4	0.0054	0.878
15	CL24	CL38	9	0.0057	0.872
14	CL39	CL19	12	0.0062	0.866
13	CL21	CL22	13	0.0068	0.859
12	CL16	CL25	11	0.0081	0.851
11	CL44	CL20	6	0.0082	0.843
10	CL36	CL32	22	0.0091	0.834
9	CL18	CL29	23	0.0098	0.824
8	CL17	CL37	16	0.0143	0.810
7	CL11	CL15	15	0.0155	0.795
6	CL13	CL14	25	0.0160	0.779
5	CL12	CL8	27	0.0269	0.752
4	CL10	CL9	45	0.0637	0.688
3	CL5	CL6	52	0.0902	0.598
2	CL4	CL7	60	0.1390	0.459
1	CL2	CL3	112	0.4589	0.000

(2) The interpretation of results

Part I

The default statistics include means, standard deviation, skewness, kurtosis, and coefficient peaks of 20 variables.

Part II

The eigenvalues of the covariance matrix, difference values between two vertical adjacent eigenvalues, variance ratio, and the cumulative variance ratio.

The square root of the standard deviation of all the samples is 1.415481, indicating a small variability within all samples. The distance between observations (genes) is 8.952286, indicating that samples (i.e., variables) is at a remote distance.

Part III

Cluster processing

Looking at the category number, 112 observations (genes) have been merged 111 times. According to the distance, the first merging (first cluster) is gene 9 and gene 16 clustered into the first category, because the standardized Euclidean distance is the smallest between them, only 0.0001. By analogy with the process, the last is the second category CL2 and the third category CL3 clustered into one category. Figure 33.1 shows that to divide the 112 genes into five major categories were more appropriate.

33.4.2 *K-means clustering*

K-means clustering (KMC) is a partition clustering method and will not create a systematic pedigree dendrogram. It can rapidly classify and suits enormous data sample clustering. Practically, the method applied must under a certain condition which is the amount of categories that has been known before use. Therefore, without any priori knowledge, researcher should try several values of K and then decide which value is the best based on the clustering results. Otherwise apply hierarchical clustering to find out how many clusters and conduct K -means clustering eventually. Class initialization at the first step is random so that different initializations could cause different clusters which are not easy to explain.

Example 33.4 Try using K -means clustering method to rapidly cluster the 112 genes based on the data from Example 33.1.

Solution The purpose of study is the same as Example 33.3. Here we use K -means clustering.

K -means clustering chooses the initial condensation point (cluster center), classifying every sample according to Euclidean Distance Coefficient. The initial condensation point is replaced with each class' center of gravity by iterative method. And samples are ranged till classified categories are

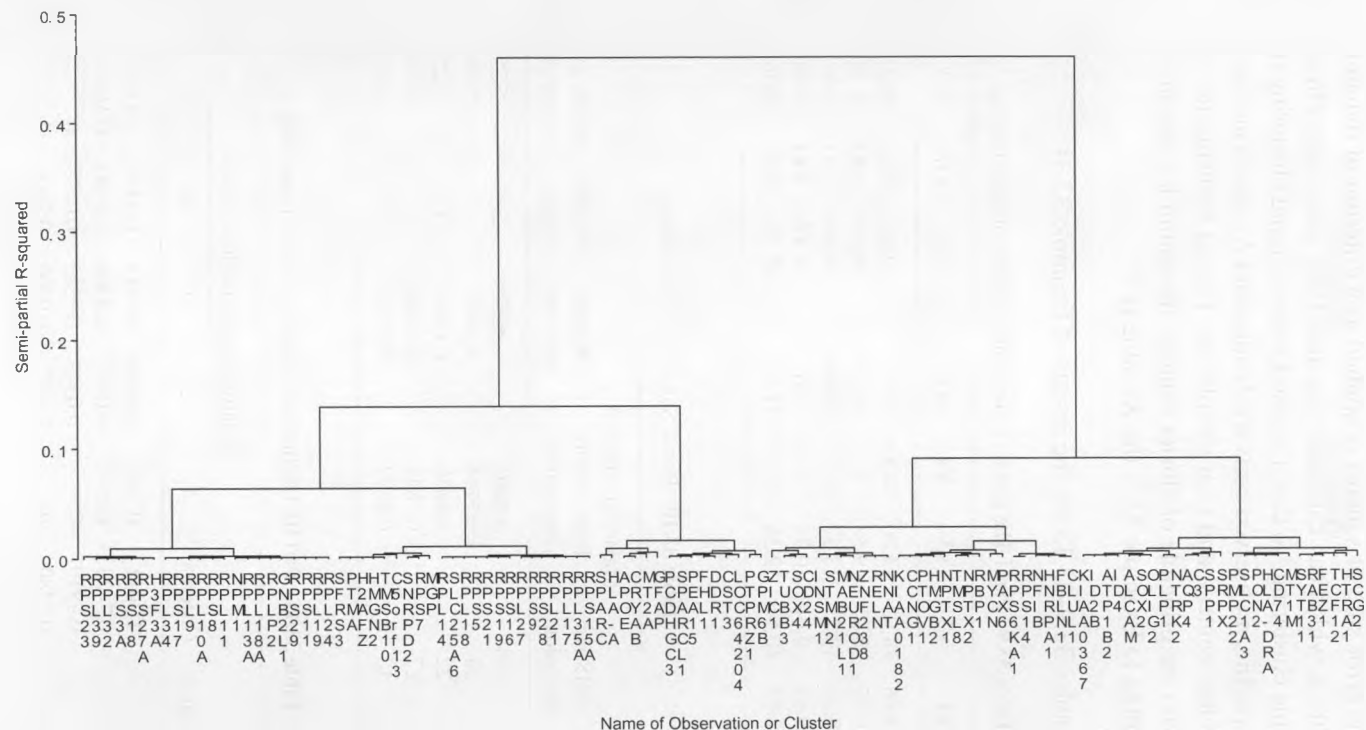


Fig. 33.1 Hierarchical clustering.

stable. The error sum of squares is applied as a criterion of dynamic clustering method, which can calculate in a short time and is very effective for large sample clustering, so that it is also known as rapid clustering method. Be more specific, n observed units are divided into K classes and determine K initial class centers, and then samples are ranged by using the iterative method with the principle of cluster centers' minimum Euclidean.

According to Example 33.3, the K -value is 5.

Output (Tables 33.9–33.13 are the results of Example 33.3)

Table 33.9 Part I The initial cluster centres of the initial clustering.

Cluster	x1	x2	x3	X4	x5	...	x17	x18	x19	x20
1	14.66	14.02	14.80	14.84	14.82	...	14.66	14.02	14.80	14.84
2	9.32	8.95	9.38	9.25	9.25	...	9.32	8.95	9.38	9.25
3	9.46	13.21	13.57	13.22	12.66	...	9.46	13.21	13.57	13.22
4	9.47	9.61	10.78	9.34	9.61	...	9.47	9.61	10.78	9.34
5	12.95	10.54	12.36	10.95	11.13	...	12.95	10.54	12.36	10.95

Table 33.10 Part II Summary of clustering.

Cluster frequency	RMS Std deviation	Maximum distance		Radius exceeded	Nearest cluster	Distance between cluster centroids
		from seed to observation				
1	26	0.5292		4.9657	3	5.2514
2	24	0.7714		4.8098	4	5.6995
3	22	0.6850		6.1436	1	5.2514
4	27	0.7843		5.1499	2	5.6995
5	13	0.8392		5.3769	4	6.8952

Table 33.11 Part III The historical iterative cluster processing.

Iteration	Criterion	Relative change in cluster seeds				
		1	2	3	4	5
1	1.3011	0.5635	0.4684	0.6614	0.5450	0.4300
2	0.7423	0.0133	0.0447	0.0506	0.0400	0.1652
3	0.7310	0	0.0309	0.0263	0.0177	0
4	0.7280	0	0.107	0.0264	0.0806	0

Table 33.12 Part IV Cluster statistics.

Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
x1	1.67843	0.73099	0.817155	4.469127
x2	1.37159	0.71129	0.740756	2.857364
x3	1.37762	0.65240	0.783811	3.625592
x4	1.63788	0.66578	0.840722	5.278325
x5	1.47203	0.55568	0.862634	6.279837
x6	1.59134	0.66905	0.829606	4.868760
x7	1.57422	0.64009	0.840626	5.274539
x8	1.38615	1.09706	0.396192	0.656154
x9	1.71731	0.77297	0.804706	4.120487
x10	1.55259	0.61615	0.848182	5.586813
x11	1.32587	0.77694	0.668998	2.021127
x12	1.19493	0.65422	0.711048	2.460783
x13	1.17325	0.62581	0.725745	2.646237
x14	1.16305	0.74230	0.607329	1.546663
x15	1.26821	0.63940	0.754970	3.081133
x16	1.20563	0.62195	0.743467	2.898131
x17	1.41017	0.71747	0.750467	3.007485
x18	1.50161	0.92573	0.633637	1.729531
x19	1.25456	0.65641	0.736108	2.789423
x20	1.24101	0.67021	0.718852	2.556841
OVER-ALL	1.41548	0.71686	0.752760	3.044650

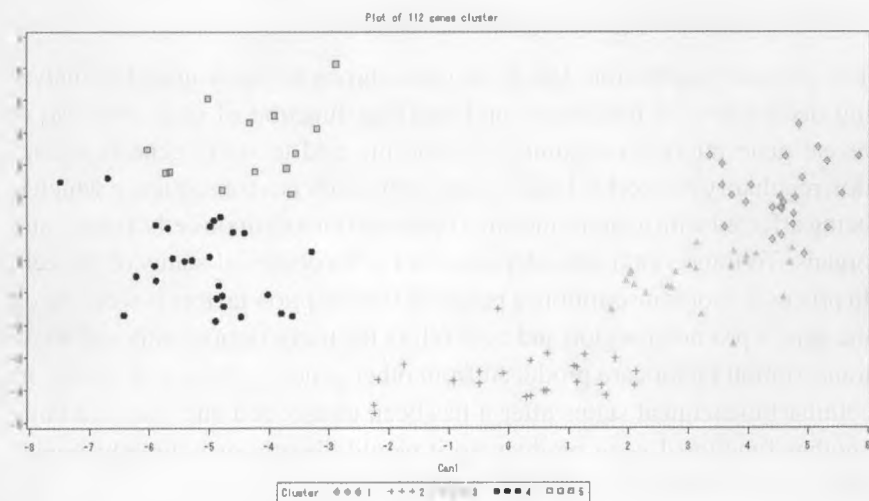
**Fig. 33.2** Five types clustering graphic illustration.

Table 33.13 Genes in different clusters.

Cluster	Number of genes	Name of genes
1	26	RPS23, RPL39, RPL14, SLC25A6, RPL34, H3F3A, RPS17, NPM1, GNB2L1, RPS3A, RPS27A, RPL9, RPL10A, RPS5, RPS29, RPL32, RPS11, RPL13A, RPL19, RPS8, RPS18, RPL11, RPL18A, RPLP2, RPL24, RPS21
2	24	KIAA0182, TMSL8, RPS6KA1, RBP1, MYCN, RPS14, INSM1, PAX6, STMN2, NPTX2, MAB21L1, ZNF238, TUBB3, NPTX1, NFIB, PTOV1, RELN, FBLN1, CCNG1, HNRPA1, NEUROD1, HMGB2, NNAT, CLUL1
3	22	SFRS3, ZIC1, RPL18, RPL27, H2AFZ, RPS16, RPS19, RPS9, SNRPD2, RPL21, C5orf13, RPS28, HMGN2, RPL35A, MGP, RPL17, RPS15A, SOX4, RPS7, CD24, PTMA, TMSB10
4	27	KIAA0367, PMP22, SLC1A3, SYT11, AQP4, HTRA1, C3, ALCAM, PON2, SCG2, PLP1, TCF12, PTPRZ1, SPP1, ID4, GPM6B, OLIG2, FEZ1, SRPX, HLA-DRA, NTRK2, ID2B, MT1M, SOX2, CD74, ATP1B2, RAB31
5	13	SPARC, PCDHGC3, APOE, CST3, CRYAB, FHL1, SPARCL1, PEA15, HLA-A, DDR1, MT2A, GFAP, LOC64204

33.5 Analysis of Gene Regulatory Networks

33.5.1 Analysis of gene regulatory networks

The obtained expression data from gene chip is not only used for analyzing disciplinary of time-space and studying function of gene, but also to reveal gene internal restraining relationship and to study gene transcription regulatory networks. In fact, gene expression is a subsequence which is being affected with genetic and environmental factors upon cells, tissues and organs. To transcript a gene depends on the biochemical status of the cell. In processing of transcription, a bunch of transcription factors is working on the gene's promoter region and controlling the transcription, although these transcription factors are produced from other genes. A gene will change its cellular biochemical status after it has been transcribed and translated into another functional gene product, so it would directly or indirectly impact other genes' expression, even its own expression.

For more than one gene, if gene expressions keep changing, the cellular biochemical status will also change. Generally, a gene expression is

impacted by other genes, and this gene also influences others, such an internal impact and internal restraining relationship has formed complicated gene expression regulatory networks. From a systematic viewpoint, a cell looks like a complex dynamic system where each gene is seemed as a systematic variable, and variables interact with another.

The purpose of the transcriptional regulatory network analysis is to establish a mathematical model of regulatory network to analyze the interaction between genes. Now many laboratories and researchers combine bio-chip technology and information technology to explore gene regulatory networks, and made some effective results. Here we briefly report the progress of gene regulatory networks associated with the mathematical model and its application in recent years to apply. Weight matrices were first applied to gene regulatory networks approach. Weaver *et al.* showed the influence between each gene by a weighted matrix, Reinitz and Sharp constructed *Drosophila* gene regulatory networks using the weighting matrix models, in order to describe the mechanism of the fruit fly gene in the *Drosophila* gene stripe formation, and find genes playing an important role in the section of *Drosophila*; Boolean algebra model, a Boolean network contains n nodes (representing genes), respectively, in the suppression or expression status (i.e., 0 and 1 states). Network is a dynamic process of Boolean functions of n determined by the state, determined by a function of each node. Therefore, the next state of the network can be decided by all nodes in the input and the function of the node. Thieffry and Thomas studied the logic of Boolean model about gene regulatory network, a detailed analysis of sea urchin *Strongylocentrotus Purpuratus* gene *Endo116*, examined how the level of gene transcription of the gene regulatory networks for accurate description of the logic. They describe the gene cis-regulatory system based on Boolean theory, and it can simulate the *Endo116* expression of transcription in the given conditions; Chen proposed a differential equation model of gene regulatory networks. They have done a lot of assumptions, such as linear transfer function, gene networks have some stability, and they use the Fourier transform technique about stability of the system to determine the various parameters; sharing information associated with the network model, Butte, etc. sharing information between two genes by calculating each, and analyze the expression data from yeast microarray. They firstly calculated shared information among all genes according to the

experimental data of gene expression, and define a share information threshold; above the threshold are considered among the genes on the biological significance of association, the connection of these genes together, sharing information to build the associated network. Correlation coefficient model is a classical method for potential biological causal relationship. Although the correlation analysis cannot provide an actual basis between the causal relationship, it provides us with a hypothesis, which might be tested by other methods. According to the principles of gene regulation, if gene *A* and gene *B* have a high correlation, it might mean: gene *A* regulates gene *B*; gene *B* regulates gene *A*, gene *A* and gene *B* are the third co-regulation of gene *C*, random-control relationship. Of course, all these regulatory relationships may be indirect in nature.

33.5.2 *Established relevant analysis model by gene regulatory networks*

Note the linear correlation coefficient formula

$$r_{ij} = \frac{\sum_{k=1}^n (x_{ik} - \bar{X}_i)(x_{jk} - \bar{X}_j)}{\sqrt{\sum_{k=1}^n (x_{ik} - \bar{X}_i)^2} \sqrt{\sum_{k=1}^n (x_{jk} - \bar{X}_j)^2}}, \quad (33.3)$$

where x_{ik} is an expression level of gene *i* under the experimental condition *k*, and \bar{X}_i is an average expression level of gene *i* under *n* experimental conditions. Between the expression patterns, a positive correlation is associated with the Euclidean distance principle, but is not sensitive to the translation transformation, additionally, the Euclidean distance analysis cannot describe negative correlation, but it implies that there may have a strong connection between the two genes.

Example 33.5 Carry out a correlation analysis over 112 genes (Example 33.1) and examine the relationship among the classes and the relationship among the genes within each class.

Solution The correlation between gene clusters: Table 33.14 gives the correlation coefficient matrix of five categories, where the coefficient represents the correlation direction and strength level among gene clusters.

Correlation within a gene cluster: Tables 33.15 and 33.16 list the correlation matrix of genes for Class 1 and Class 5 only; in some other conditions,

Table 33.14 The correlation coefficient matrix of five classes.

Classes	Class 1	Class 2	Class 3	Class 4	Class 5
Class 1	1	0.70	0.96	-0.79	-0.78
Class 2	0.70	1	0.83	-0.87	-0.86
Class 3	0.96	0.83	1	-0.87	-0.86
Class 4	-0.79	-0.87	-0.87	1	0.96
Class 5	-0.78	-0.86	-0.86	0.96	1

Table 33.15 The related matrix of genes within Class 1.

Genes	RPS23	RPL39	RPL14	...	DDR1	MT2A	GFAP	LOC64204
RPS23	1	0.90	0.90		0.92	0.91	0.91	0.94
RPL39	0.90	1	0.90		0.92	0.90	0.91	0.90
RPL14	0.90	0.90	1		0.93	0.91	0.91	0.94
...				...				
DDR1	0.92	0.92	0.93		1	0.94	0.91	0.96
MT2A	0.91	0.90	0.91		0.94	1	0.88	0.95
GFAP	0.91	0.91	0.91		0.91	0.88	1	0.92
LOC64204	0.94	0.90	0.94		0.96	0.95	0.92	1

Table 33.16 The related matrix of Genes within Class 5.

Genes	SPARC	PCDHGC3	APOE	...	MT2A	GFAP	LOC64204
SPARC	1	0.45	0.44		0.54	0.28	0.78
PCDHGC3	0.45	1	0.67		0.62	0.76	0.48
APOE	0.44	0.67	1		0.76	0.86	0.38
...				...			
DDR1	0.43	0.88	0.78		0.59	0.83	0.38
MT2A	0.54	0.62	0.76		1	0.62	0.60
GFAP	0.28	0.76	0.86		0.62	1	0.34
LOC64204	0.78	0.48	0.38		0.60	0.34	1

the matrix should be similar. The coefficients are all positive indicating that correlativity within the gene cluster is of isotropic attribute, and the value shows relevant strength. With the method, classify genes into different categories first, and then study the correlation between individual genes.

Table 33.17 The related matrix of Class 1 genes and Class 5 genes.

Gene	SPARC	PCDHGC3	APOE	CST3	...	DDR1	MT2A	GFAP	LOC64204
RPS23	−0.54	−0.72	−0.63	−0.48		−0.77	−0.42	−0.70	−0.36
RPL39	−0.48	−0.63	−0.69	−0.48		−0.69	−0.41	−0.67	−0.25
RPL14	−0.43	−0.75	−0.61	−0.53		−0.76	−0.33	−0.69	−0.28
SLC25A6	−0.38	−0.72	−0.58	−0.39		−0.71	−0.46	−0.62	−0.22
RPL34	−0.52	−0.58	−0.49	−0.39		−0.57	−0.26	−0.53	−0.40
...									
RPL11	−0.46	−0.80	−0.63	−0.42		−0.77	−0.39	−0.67	−0.35
RPL18A	−0.42	−0.75	−0.63	−0.54		−0.75	−0.49	−0.70	−0.32
RPLP2	−0.45	−0.73	−0.66	−0.61		−0.70	−0.48	−0.76	−0.35
RPL24	−0.53	−0.70	−0.63	−0.59		−0.77	−0.43	−0.73	−0.40
RPS21	−0.48	−0.84	−0.69	−0.58		−0.83	−0.52	−0.79	−0.38

It can condense the matrix size, easier to understand and find out the real relationship of biological regulation.

The biological explanation: Five types clustering results of correlation coefficient matrix shows Class 1 positively correlated with Class 2 and Class 3 and negatively correlated with both Class 4 and Class 5, and Class 4 and Class 5 were a positively correlated.

The related matrix of Class 1 genes and Class 5 genes indicated that genes were strongly positively correlated with each other in Class 1, and the overwhelming majority of the correlation coefficient is above 0.9. In Class 5, genes were mutually moderately positively correlated, the correlation coefficient mainly distributed from 0.5 to 0.8. Table 33.17 shows that the genes in Class 1 were moderately negatively correlated with those in Class 5. From clustering analysis, it shows that the relationship between two types of genes was well extracted and can provide clues for in-depth study.

33.6 Computerized Experiments

Experiment 33.1 The format of gene expression data Select two types of tissues (medulloblastoma and malignant glioma) of total 20 samples from 112 differentially expressed genes which are from Example 33.1, and list all data with SAS for further study.

Experiment 33.2 SAS programs of differential gene expression screening Explore expressed different Notch1 gene in three degree

Program 33.1 The format of gene expression data.

Line	Program	Line	Program
01	data PRG33_1;	04	proc print DATA=PRG33_1;
02	Infile 'E:\Chapter33\TXT\ genes.txt';	05	Run ;
03	Input name \$ x1-x20;		

Program 33.2 The result of differentially expressed genes analysis.

Line	Program	Line	Program
01	data PRG33_2;	16	0.505 0.127 0.057
02	do grp= 1 to 3 ;	17	0.122 0.031
03	Input Notch1 @@;	18	0.123 0.069
04	if Notch1 ne . then output;	19	0.051
05	end;	20	0.038
06	Cards;	21	0.059
07	0.552 0.109 0.058	22	0.043
08	0.513 0.101 0.046	23	;
09	0.451 0.116 0.056	24	proc glm data= PRG33_2;
10	0.451 0.125 0.052	25	class grp;
11	0.559 0.127 0.066	26	model Notch1=grp;
12	0.619 0.122 0.042	27	means grp;
13	0.502 0.112 0.042	27	run ;
14	0.527 0.102 0.049	28	quit ;
15	0.563 0.104 0.045		

varicocele testicular tissue from 38 varicoele patients, data from Example 33.2.

Experiment 33.3 Gene hierarchical clustering method Based on Example 33.3, apply hierarchical clustering to 112 differentially expressed genes which are from Program 33.1. SAS program details every step of computation below.

Experiment 33.4 Gene *k*-means clustering Use *k*-means clustering method for the rapid clustering with the 112 genes in the cases of Program 33.1 based on the data of Example 33.4.

Experiment 33.5 Method of gene networks constructional correlation coefficient From Example 33.5, construct gene networks with 112 genes (from Example 33.1) by correlation coefficient methods.

Program 33.3 Gene hierarchical clustering method.

Line	Program	Line	Program
01	Data PRG33_3;	06	id name;
02	Infile 'E:\Chapter33\TXT\genes.txt';	07	run ;
03	Input name\$ x1-x20;	08	proc print data=tree0;
04	Proc cluster data= PRG33_3 simple	09	run ;
	method=ward outtree=tree0;	10	proc tree data=tree0;
05	var x1-x20;	11	run ;

Program 33.4 Gene *k*-means clustering.

Line	Program	Line	Program
01	data PRG33_4;	16	run ;
02	Infile 'E:\Chapter33\TXT\genes.txt';	17	legend1 cframe=ligr
03	Input name\$ x1-x20 ;		cborder=black
04	proc fastclus data= PRG33_4 maxc=5	18	position=center
	maxiter=4 out=fac;		value=(justify=center);
05	var x1-x20;	19	axis1 label=(angle=90 rotate=0)
06	id name;		minor=none;
07	proc sort ;	20	axis2 minor=none;
08	by cluster;	21	proc gplot data=Can:
09	proc print data=fac;	22	plot Can2*Can1=Cluster/frame
10	var name cluster;		cframe=ligr
11	run ;	23	legend=legend1 vaxis=axis1
12	proc candisc anova out=can;		haxis=axis2;
14	var x1-x20;	24	title2 'Plot of Canonical Variables
15	title2 'Canonical Discriminant		Identified by Cluster';
	Analysis of gene Clusters';	25	run ;
14	var x1-x20;		

33.7 Summary

This chapter introduced several major statistical methods in bioinformatics. The sequence comparison method of BLAST, FASTA and Smith–Waterman not only include probability theory, but also ranged over the statistical methods of hypothesis test. Due to the rapid development of biological database, faster and accuracy statistical methods are needed. Controlling multiple testing error rates (false positive rate) is the key of differential gene expression screening, and must guarantee a high screening efficiency. Methods like SAM, two sample *t*-test, Bonferroni correction and BH are very commonly

Program 33.5 The method of gene networks constructional correlation coefficient.

Line	Program	Line	Program
01	data PRG33_5;	18	where _name_ not in
02	Infile 'E:\Chapter33\TXT\genes.txt';		("CLUSTER", "DISTANCE");
03	Input name\$ x1-x20;	19	run ;
04	proc fastclus data= PRG33_6	20	/*Genes in Cluster 1 and 5*/
	maxc=5 maxiter=4 out=fac		proc corr data=tfac;
	mean=cluster;	21	Var RPS23 – RPS21 SPARC –
05	var x1-x20;		LOC64204;
06	id name;	22	Run ;
07	proc transpose data=cluster	23	/*Genes in Cluster 1*/
	out=tclus; var x1-x20;	24	proc corr data=tfac;
08	/*Correlation Coefficients Between	25	Var RPS23 – RPS21;
	Clusters*/	26	Run ;
09	proc corr data=tclus out=rtclus ;	27	/*Genes in Cluster 5*/
10	run ;	28	proc corr data=tfac;
11	/*Correlation Coefficients Between	29	Var SPARC – LOC64204;
	Genes*/	30	Run ;
12	proc sort data=fac;	31	/*Genes in Cluster 1 VS 5*/
13	by cluster;	32	proc corr data=tfac;
14	proc transpose data=fac out=tfac;	33	with RPS23 – RPS21;
15	id name;	34	Var SPARC – LOC64204;
16	data tfac;	35	Run ;
17	set tfac;		

used for microarray data in selection of differentially expressed genes. However, some new statistical thinking occurred, it avoided a restriction which is pairwise comparison. Gene chip is an important biological analysis technology, that makes contribution in genetic function identification, tumor disease diagnosis, pathogenic mechanism analysis and drug design. At present, mature methods are available in some key processes, including chip design, image acquisition, data processing and analysis, etc. Data clustering analysis is the important part of gene microarray data analysis, which is widely applied in disease diagnosis. Clustering methods mainly included unsupervised clustering and supervised clustering. Gene chip technologies are widely applied to gene regulation networks studies that contribute to elucidate some disease mechanism and are currently a hot topic in bioinformatics. Software of gene chip data analysis includes ScanAlyze for image

acquisition, Cluster for clustering analysis, TreeView for displaying the results of clustering, etc.

However, the current bioinformatics comes up against many difficulties and challenges. The structure, mechanism and function of biology are awaiting experimental verification, technical analysis methods and theories still need to be improved and updated. Therefore, some methods or implementation in this chapter are imperfect due to the restriction of the software and method itself. Refinement will be focused in further study.

33.8 Practice and Experiment

1. Can you determine the possible virus species by the sequence of the SARS virus?
2. What is the significance of the sequence comparability search?
3. Describe the relationship between similarity and homology of the biological sequences.
4. To realize the gene chip application prospects, present situation of application and the mainly restricting factors during gene chip development through the literature review.
5. Observe the data format and the results analysis from a cancer gene chip database.

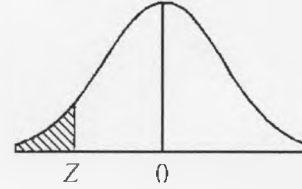
(2nd edn. Dong Yi)

Appendix II

Statistical Tables

Table 1. Distribution function of standard normal distribution.

$$\Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^Z e^{-\frac{x^2}{2}} dx \quad (Z \leq 0)$$



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	Z
-0.0	0.500 0	0.496 0	0.492 0	0.488 0	0.484 0	0.480 1	0.476 1	0.472 1	0.468 1	0.464 1	-0.0
-0.1	0.460 2	0.456 2	0.452 2	0.448 3	0.444 3	0.440 4	0.436 4	0.432 5	0.428 6	0.424 7	-0.1
-0.2	0.420 7	0.416 8	0.412 9	0.409 0	0.405 2	0.401 8	0.397 4	0.393 6	0.389 7	0.385 9	-0.2
-0.3	0.382 1	0.378 3	0.374 5	0.370 7	0.366 9	0.363 2	0.359 4	0.355 7	0.352 0	0.348 3	-0.3
-0.4	0.344 6	0.340 9	0.337 2	0.333 6	0.330 0	0.326 4	0.322 8	0.319 2	0.315 6	0.312 1	-0.4
-0.5	0.308 5	0.305 0	0.301 5	0.298 1	0.294 6	0.291 2	0.287 7	0.284 3	0.281 0	0.277 6	-0.5
-0.6	0.274 3	0.270 9	0.267 6	0.264 3	0.261 1	0.257 8	0.254 6	0.251 4	0.248 3	0.245 1	-0.6
-0.7	0.242 0	0.238 9	0.235 8	0.232 7	0.229 7	0.226 6	0.223 6	0.220 6	0.217 7	0.214 8	-0.7
-0.8	0.211 9	0.209 0	0.206 1	0.203 3	0.200 5	0.197 7	0.194 9	0.192 2	0.189 4	0.186 7	-0.8
-0.9	0.184 1	0.181 4	0.178 8	0.176 2	0.173 6	0.171 1	0.168 5	0.166 0	0.163 5	0.161 1	-0.9
-1.0	0.158 7	0.156 2	0.153 9	0.151 5	0.149 2	0.146 9	0.144 6	0.142 3	0.140 1	0.137 9	-1.0
-1.1	0.135 7	0.133 5	0.131 4	0.129 2	0.127 1	0.125 1	0.123 0	0.121 0	0.119 0	0.117 0	-1.1
-1.2	0.115 1	0.113 1	0.111 2	0.109 3	0.107 5	0.105 6	0.103 8	0.102 0	0.100 3	0.098 53	-1.2
-1.3	0.096 80	0.095 10	0.093 42	0.091 76	0.090 12	0.088 51	0.086 91	0.085 34	0.083 79	0.082 26	-1.3
-1.4	0.080 76	0.079 27	0.077 80	0.076 36	0.074 93	0.073 53	0.072 15	0.070 78	0.069 44	0.068 11	-1.4

Table 1. (Continued)

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	Z
-1.5	0.066 81	0.065 52	0.064 26	0.063 01	0.061 78	0.060 57	0.059 38	0.058 21	0.057 05	0.055 92	-1.5
-1.6	0.054 80	0.053 70	0.052 62	0.051 55	0.050 50	0.049 47	0.048 46	0.047 46	0.046 48	0.045 51	-1.6
-1.7	0.044 57	0.043 63	0.042 72	0.041 82	0.040 93	0.040 06	0.039 20	0.038 36	0.037 54	0.036 73	-1.7
-1.8	0.035 93	0.035 15	0.034 38	0.033 62	0.032 88	0.032 16	0.031 44	0.030 74	0.030 05	0.029 38	-1.8
-1.9	0.028 72	0.028 07	0.027 43	0.026 80	0.026 19	0.025 59	0.025 00	0.024 42	0.023 85	0.023 30	-1.9
-2.0	0.022 75	0.022 22	0.021 69	0.021 18	0.020 68	0.020 18	0.019 70	0.019 23	0.018 76	0.018 31	-2.0
-2.1	0.017 86	0.017 43	0.017 00	0.016 59	0.016 18	0.015 78	0.015 39	0.015 00	0.014 63	0.014 26	-2.1
-2.2	0.013 90	0.013 55	0.013 21	0.012 87	0.012 55	0.012 22	0.011 91	0.011 60	0.011 30	0.011 01	-2.2
-2.3	0.010 72	0.010 44	0.010 17	0.009 903	0.009 642	0.009 387	0.009 137	0.008 894	0.008 656	0.008 424	-2.3
-2.4	0.008 198	0.007 976	0.007 760	0.007 549	0.007 344	0.007 143	0.006 947	0.006 756	0.006 569	0.006 387	-2.4
-2.5	0.006 210	0.006 037	0.005 868	0.005 703	0.005 543	0.005 386	0.005 234	0.005 085	0.004 940	0.004 799	-2.5
-2.6	0.004 661	0.004 527	0.004 396	0.004 269	0.004 145	0.004 025	0.003 907	0.003 793	0.003 681	0.003 573	-2.6
-2.7	0.003 467	0.003 364	0.003 264	0.003 167	0.003 072	0.002 980	0.002 890	0.002 803	0.002 718	0.002 635	-2.7
-2.8	0.002 555	0.002 477	0.002 401	0.002 327	0.002 256	0.002 186	0.002 118	0.002 052	0.001 938	0.001 826	-2.8
-2.9	0.001 866	0.001 807	0.001 750	0.001 695	0.001 641	0.001 589	0.001 538	0.001 489	0.001 441	0.001 395	-2.9
-3.0	0.001 350	0.001 306	0.001 264	0.001 223	0.001 183	0.001 144	0.001 107	0.001 070	0.001 035	0.001 001	-3.0
-3.1	0.000 9676	0.000 9354	0.000 9043	0.000 8740	0.000 8447	0.000 8164	0.000 7888	0.000 7622	0.000 7364	0.000 7114	-3.1
-3.2	0.000 6871	0.000 6637	0.000 6410	0.000 6190	0.000 5976	0.000 5770	0.000 5571	0.000 5377	0.000 5190	0.000 5009	-3.2
-3.3	0.000 4834	0.000 4665	0.000 4501	0.000 4342	0.000 4189	0.000 4041	0.000 3897	0.000 3758	0.000 3624	0.000 3495	-3.3
-3.4	0.000 3369	0.000 3248	0.000 3131	0.000 3018	0.000 2909	0.000 2803	0.000 2701	0.000 2602	0.000 2507	0.000 2415	-3.4

Table 1. (Continued)

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	Z
-3.5	0.0 ³ 23 26	0.0 ³ 22 41	0.0 ³ 21 58	0.0 ³ 20 78	0.0 ³ 20 01	0.0 ³ 19 26	0.0 ³ 18 54	0.0 ³ 17 85	0.0 ³ 17 18	0.0 ³ 16 53	-3.5
-3.6	0.0 ³ 15 91	0.0 ³ 15 31	0.0 ³ 14 73	0.0 ³ 14 17	0.0 ³ 13 63	0.0 ³ 13 11	0.0 ³ 12 61	0.0 ³ 12 13	0.0 ³ 11 66	0.0 ³ 11 21	-3.6
-3.7	0.0 ³ 10 78	0.0 ³ 10 36	0.0 ⁴ 09 61	0.0 ⁴ 09 54	0.0 ⁴ 09 21	0.0 ⁴ 08 42	0.0 ⁴ 08 46	0.0 ⁴ 08 16	0.0 ⁴ 07 81	0.0 ⁴ 07 53	-3.7
-3.8	0.0 ⁴ 07 35	0.0 ⁴ 06 48	0.0 ⁴ 06 73	0.0 ⁴ 06 07	0.0 ⁴ 06 52	0.0 ⁴ 05 06	0.0 ⁴ 05 69	0.0 ⁴ 05 42	0.0 ⁴ 05 23	0.0 ⁴ 05 12	-3.8
-3.9	0.0 ⁴ 04 10	0.0 ⁴ 04 15	0.0 ⁴ 04 27	0.0 ⁴ 04 27	0.0 ⁴ 04 74	0.0 ⁴ 03 08	0.0 ⁴ 03 47	0.0 ⁴ 03 59	0.0 ⁴ 03 46	0.0 ⁴ 03 04	-3.9
-4.0	0.0 ⁴ 03 17	0.0 ⁴ 03 36	0.0 ⁴ 02 91	0.0 ⁴ 02 89	0.0 ⁴ 02 73	0.0 ⁴ 02 51	0.0 ⁴ 02 54	0.0 ⁴ 02 51	0.0 ⁴ 02 52	0.0 ⁴ 02 17	-4.0
-4.1	0.0 ⁴ 02 06	0.0 ⁴ 01 78	0.0 ⁴ 01 39	0.0 ⁴ 01 14	0.0 ⁴ 01 37	0.0 ⁴ 01 62	0.0 ⁴ 01 51	0.0 ⁴ 01 53	0.0 ⁴ 01 58	0.0 ⁴ 01 39	-4.1
-4.2	0.0 ⁴ 01 35	0.0 ⁴ 01 27	0.0 ⁴ 01 22	0.0 ⁴ 01 18	0.0 ⁴ 01 18	0.0 ⁴ 01 09	0.0 ⁴ 01 22	0.0 ⁴ 01 22	0.0 ⁴ 01 22	0.0 ⁴ 01 22	-4.2
-4.3	0.0 ⁵ 08 40	0.0 ⁵ 08 16	0.0 ⁵ 07 81	0.0 ⁵ 07 55	0.0 ⁵ 07 24	0.0 ⁵ 06 07	0.0 ⁵ 06 03	0.0 ⁵ 06 12	0.0 ⁵ 05 34	0.0 ⁵ 05 68	-4.3
-4.4	0.0 ⁵ 05 13	0.0 ⁵ 05 16	0.0 ⁵ 04 35	0.0 ⁵ 04 12	0.0 ⁵ 04 48	0.0 ⁵ 04 29	0.0 ⁵ 04 08	0.0 ⁵ 03 91	0.0 ⁵ 03 32	0.0 ⁵ 03 61	-4.4
-4.5	0.0 ⁵ 03 98	0.0 ⁵ 03 41	0.0 ⁵ 03 02	0.0 ⁵ 02 49	0.0 ⁵ 02 13	0.0 ⁵ 02 82	0.0 ⁵ 02 58	0.0 ⁵ 02 43	0.0 ⁵ 02 25	0.0 ⁵ 02 16	-4.5
-4.6	0.0 ⁵ 02 12	0.0 ⁵ 02 18	0.0 ⁵ 01 19	0.0 ⁵ 01 28	0.0 ⁵ 01 42	0.0 ⁵ 01 60	0.0 ⁵ 01 51	0.0 ⁵ 01 56	0.0 ⁵ 01 34	0.0 ⁵ 01 66	-4.6
-4.7	0.0 ⁵ 01 01	0.0 ⁵ 01 39	0.0 ⁵ 01 17	0.0 ⁵ 01 23	0.0 ⁵ 01 09	0.0 ⁵ 01 17	0.0 ⁵ 01 08	0.0 ⁵ 01 11	0.0 ⁵ 00 87	0.0 ⁵ 00 39	-4.7
-4.8	0.0 ⁶ 07 33	0.0 ⁶ 07 47	0.0 ⁶ 07 18	0.0 ⁶ 06 27	0.0 ⁶ 06 42	0.0 ⁶ 06 17	0.0 ⁶ 05 69	0.0 ⁶ 05 80	0.0 ⁶ 05 04	0.0 ⁶ 04 42	-4.8
-4.9	0.0 ⁶ 04 92	0.0 ⁶ 04 54	0.0 ⁶ 04 27	0.0 ⁶ 04 11	0.0 ⁶ 03 06	0.0 ⁶ 03 11	0.0 ⁶ 03 25	0.0 ⁶ 03 48	0.0 ⁶ 03 19	0.0 ⁶ 03 19	-4.9

Table 1. (Continued)

$$\Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^Z e^{-\frac{x^2}{2}} dx \quad (Z \geq 0)$$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	Z
0.0	0.500 0	0.504 0	0.508 0	0.512 0	0.516 0	0.519 9	0.623 9	0.527 9	0.531 9	0.535 9	0.0
0.1	0.539 8	0.543 8	0.547 8	0.551 7	0.555 7	0.559 6	0.563 6	0.567 5	0.571 4	0.575 3	0.1
0.2	0.579 3	0.583 2	0.587 1	0.591 0	0.594 3	0.598 7	0.602 6	0.606 4	0.610 3	0.614 1	0.2
0.3	0.617 9	0.621 7	0.625 5	0.629 3	0.633 1	0.636 8	0.640 6	0.644 3	0.643 0	0.651 7	0.3
0.4	0.655 4	0.659 1	0.662 8	0.666 4	0.670 0	0.673 6	0.677 2	0.680 8	0.684 4	0.687 9	0.4
0.5	0.691 5	0.695 0	0.698 5	0.701 9	0.705 4	0.708 8	0.712 3	0.715 7	0.710 0	0.722 4	0.5
0.6	0.725 7	0.729 1	0.732 4	0.735 7	0.738 9	0.742 2	0.745 4	0.748 6	0.751 7	0.754 9	0.6
0.7	0.758 0	0.761 1	0.764 2	0.767 3	0.770 3	0.773 4	0.776 4	0.779 4	0.782 3	0.785 2	0.7
0.8	0.788 1	0.791 0	0.793 9	0.796 7	0.799 5	0.802 3	0.805 1	0.807 8	0.810 6	0.813 3	0.8
0.9	0.815 9	0.818 6	0.821 2	0.823 8	0.826 4	0.828 9	0.831 5	0.834 0	0.836 5	0.838 9	0.9
1.0	0.841 3	0.843 8	0.846 1	0.848 5	0.850 8	0.853 1	0.855 4	0.857 7	0.859 9	0.862 1	1.0
1.1	0.854 3	0.866 5	0.868 6	0.870 8	0.872 9	0.874 9	0.877 0	0.879 0	0.881 0	0.883 0	1.1
1.2	0.884 9	0.885 9	0.888 8	0.890 7	0.892 5	0.894 4	0.896 2	0.898 0	0.899 7	0.901 48	1.2
1.3	0.903 20	0.904 90	0.906 58	0.908 24	0.908 88	0.911 49	0.913 09	0.914 66	0.916 21	0.917 74	1.3
1.4	0.919 24	0.920 73	0.922 20	0.923 64	0.925 07	0.926 47	0.927 85	0.929 22	0.930 56	0.931 89	1.4
1.5	0.933 19	0.934 48	0.935 74	0.936 99	0.938 22	0.939 43	0.940 62	0.941 79	0.942 95	0.944 08	1.5
1.6	0.945 20	0.946 30	0.947 38	0.948 45	0.949 50	0.950 53	0.951 54	0.952 54	0.953 52	0.954 49	1.6
1.7	0.955 43	0.956 37	0.957 28	0.958 18	0.959 07	0.959 94	0.960 80	0.961 64	0.952 46	0.963 27	1.7
1.8	0.964 07	0.964 85	0.965 62	0.966 38	0.967 12	0.967 84	0.968 56	0.969 26	0.969 95	0.970 62	1.8
1.9	0.971 28	0.971 93	0.972 57	0.973 20	0.973 81	0.974 41	0.975 00	0.975 58	0.976 15	0.976 70	1.9

Table 1. (Continued)

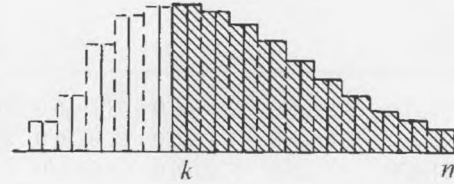
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	Z
2.0	0.977 25	0.977 78	0.978 31	0.978 82	0.979 32	0.979 82	0.980 30	0.980 77	0.981 24	0.981 69	2.0
2.1	0.982 14	0.982 57	0.983 00	0.983 41	0.983 82	0.984 22	0.984 61	0.985 00	0.985 37	0.985 74	2.1
2.2	0.986 10	0.986 45	0.986 79	0.987 13	0.987 45	0.987 78	0.988 09	0.988 40	0.988 70	0.988 99	2.2
2.3	0.989 28	0.989 56	0.989 83	0.9200 97	0.9203 68	0.9206 13	0.9208 63	0.9211 06	0.9213 44	0.9215 76	2.3
2.4	0.9218 02	0.9220 24	0.9222 40	0.9224 51	0.9226 56	0.9228 57	0.9230 53	0.9232 44	0.9234 31	0.9236 13	2.4
2.5	0.9237 90	0.9239 63	0.9241 32	0.9242 97	0.9244 57	0.9246 14	0.9247 66	0.9249 15	0.9250 60	0.9252 01	2.5
2.6	0.9253 39	0.9254 73	0.9256 04	0.9257 31	0.9258 55	0.9259 75	0.9260 93	0.9262 07	0.9263 19	0.9264 27	2.6
2.7	0.9265 33	0.9266 36	0.9267 36	0.9268 33	0.9269 28	0.9270 20	0.9271 10	0.9271 97	0.9272 82	0.9273 69	2.7
2.8	0.9274 45	0.9275 23	0.9275 99	0.9276 73	0.9277 44	0.9278 14	0.9278 82	0.9279 48	0.9280 12	0.9280 74	2.8
2.9	0.9281 34	0.9281 93	0.9282 50	0.9283 05	0.9283 59	0.9284 11	0.9284 62	0.9285 11	0.9285 59	0.9286 05	2.9
3.0	0.9286 50	0.9286 94	0.9287 35	0.9287 77	0.9288 17	0.9288 56	0.9288 93	0.9289 30	0.9289 65	0.9289 99	3.0
3.1	0.9303 24	0.9306 46	0.9309 57	0.9312 50	0.9315 53	0.9318 36	0.9321 12	0.9323 78	0.9326 36	0.9328 86	3.1
3.2	0.9331 29	0.9333 63	0.9335 90	0.9338 10	0.9340 24	0.9342 30	0.9344 29	0.9346 23	0.9348 10	0.9349 91	3.2
3.3	0.9351 66	0.9353 35	0.9354 99	0.9356 58	0.9358 11	0.9359 59	0.9361 03	0.9362 42	0.9363 76	0.9365 05	3.3
3.4	0.9366 31	0.9367 52	0.9368 69	0.9369 82	0.9370 91	0.9371 97	0.9372 99	0.9373 98	0.9374 93	0.9375 85	3.4

Table 1. (Continued)

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	Z
3.5	0.9376 74	0.9377 59	0.9378 42	0.9379 22	0.9379 99	0.9380 74	0.9381 46	0.9382 15	0.9382 82	0.9383 47	3.5
3.6	0.9384 09	0.9384 69	0.9385 27	0.9385 83	0.9386 39	0.9386 89	0.9387 39	0.9387 87	0.9388 34	0.9388 79	3.6
3.7	0.9389 22	0.9389 64	0.9390 39	0.9390 26	0.9390 99	0.9391 58	0.9392 04	0.9392 38	0.9392 59	0.9392 58	3.7
3.8	0.9427 65	0.9430 52	0.9433 27	0.9435 93	0.9438 48	0.9440 94	0.9443 31	0.9445 58	0.9447 77	0.9449 88	3.8
3.9	0.9451 90	0.9453 85	0.9455 73	0.9457 53	0.9459 26	0.9460 92	0.9462 53	0.9464 06	0.9465 54	0.9466 96	3.9
4.0	0.9468 33	0.9469 64	0.9470 90	0.9472 11	0.9473 27	0.9474 39	0.9475 46	0.9476 49	0.9477 48	0.9478 43	4.0
4.1	0.9479 34	0.9480 22	0.9481 06	0.9481 86	0.9482 63	0.9483 38	0.9484 09	0.9484 77	0.9485 42	0.9486 05	4.1
4.2	0.9486 65	0.9487 23	0.9487 78	0.9488 32	0.9488 82	0.9489 31	0.9489 78	0.9490 26	0.9490 55	0.9490 66	4.2
4.3	0.9514 60	0.9518 37	0.9521 99	0.9525 45	0.9528 76	0.9531 93	0.9534 97	0.9537 88	0.9540 66	0.9543 32	4.3
4.4	0.9545 87	0.9548 31	0.9550 65	0.9552 82	0.9555 02	0.9557 06	0.9559 02	0.9560 89	0.9562 68	0.9564 39	4.4
4.5	0.9566 02	0.9567 59	0.9569 08	0.9570 51	0.9571 87	0.9573 18	0.9574 42	0.9575 61	0.9576 75	0.9577 84	4.5
4.6	0.9578 88	0.9579 87	0.9580 81	0.9581 72	0.9582 58	0.9583 40	0.9584 19	0.9584 91	0.9585 66	0.9586 34	4.6
4.7	0.9586 99	0.9587 61	0.9588 21	0.9588 77	0.9589 31	0.9589 83	0.9590 32	0.9590 79	0.9591 23	0.9591 61	4.7
4.8	0.9620 67	0.9624 53	0.9628 22	0.9631 73	0.9635 08	0.9638 27	0.9641 31	0.9644 20	0.9646 96	0.9649 58	4.8
4.9	0.9652 08	0.9654 46	0.9656 73	0.9658 89	0.9660 94	0.9662 89	0.9664 75	0.9666 52	0.9668 21	0.9669 31	4.9

Table 2. The upper probability of binomial distribution.

$$Q(n, k, \pi) = \sum_{i=k}^n \binom{n}{i} \pi^i (1 - \pi)^{n-i}$$



$\pi \backslash k$		0.01	0.02	0.04	0.06	0.08	0.1	0.2	0.3	0.4	0.5	$\pi \backslash k$	n
5	5			0.000 00	0.000 00	0.000 00	0.000 01	0.000 32	0.002 43	0.010 24	0.031 25	5	5
	4	0.000 00	0.000 00	0.000 01	0.000 06	0.000 19	0.000 46	0.006 72	0.030 78	0.087 04	0.187 50	4	
	3	0.000 01	0.000 08	0.000 60	0.001 97	0.004 53	0.008 56	0.057 92	0.163 08	0.317 44	0.500 00	3	
	2	0.000 98	0.003 84	0.014 76	0.031 87	0.054 36	0.081 46	0.262 72	0.471 78	0.663 04	0.812 50	2	
	1	0.049 01	0.096 08	0.184 63	0.266 10	0.340 92	0.409 51	0.672 32	0.831 93	0.922 24	0.968 75	1	
10	10								0.000 01	0.000 10	0.000 98	10	10
	9							0.000 00	0.000 14	0.001 68	0.010 74	9	
	8						0.000 00	0.000 08	0.001 59	0.012 29	0.054 69	8	
	7				0.000 00	0.000 00	0.000 01	0.000 86	0.010 59	0.054 76	0.171 88	7	
	6			0.000 00	0.000 01	0.000 04	0.000 15	0.006 37	0.047 35	0.166 24	0.376 95	6	
	5		0.000 00	0.000 02	0.000 15	0.000 59	0.001 63	0.032 79	0.150 27	0.366 90	0.623 05	5	
	4	0.000 00	0.000 03	0.000 44	0.002 03	0.005 80	0.012 80	0.120 87	0.350 39	0.617 72	0.828 13	4	
	3	0.000 11	0.000 86	0.006 21	0.018 84	0.040 08	0.070 19	0.322 20	0.617 22	0.832 71	0.945 31	3	
	2	0.004 27	0.016 18	0.058 15	0.117 59	0.187 88	0.263 90	0.624 19	0.350 69	0.953 64	0.989 26	2	
	1	0.095 62	0.182 93	0.335 17	0.461 38	0.565 61	0.651 32	0.892 63	0.971 75	0.993 95	0.999 02	1	

Table 2. (Continued)

n	$k \backslash \pi$	0.01	0.02	0.04	0.06	0.08	0.1	0.2	0.3	0.4	0.5	$\pi \backslash k$	n
15	15									0.000 00	0.000 03	15	15
	14								0.000 00	0.000 03	0.000 49	14	
	13								0.000 01	0.000 28	0.003 69	13	
	12							0.000 00	0.000 09	0.001 93	0.017 58	12	
	11							0.000 01	0.000 67	0.009 35	0.059 23	11	
	10							0.000 11	0.003 65	0.033 83	0.150 88	10	
	9					0.000 00	0.000 00	0.000 79	0.015 24	0.095 05	0.303 62	9	
	8				0.000 00	0.000 01	0.000 03	0.004 24	0.050 01	0.213 10	0.500 00	8	
	7			0.000 00	0.000 01	0.000 08	0.000 31	0.018 06	0.131 14	0.390 19	0.696 38	7	
	6		0.000 00	0.000 01	0.000 15	0.000 70	0.002 25	0.061 05	0.278 38	0.596 78	0.849 12	6	
	5	0.000 00	0.000 01	0.000 22	0.001 40	0.004 97	0.012 72	0.164 23	0.484 51	0.782 72	0.940 77	5	
	4	0.000 01	0.000 18	0.002 45	0.010 36	0.027 31	0.055 56	0.351 84	0.703 13	0.909 50	0.982 42	4	
	3	0.000 42	0.003 04	0.020 29	0.057 13	0.112 97	0.184 06	0.601 98	0.873 17	0.972 89	0.996 31	3	
	2	0.009 63	0.035 34	0.119 11	0.226 24	0.340 27	0.450 96	0.832 87	0.964 73	0.994 83	0.999 51	2	
	1	0.139 94	0.261 43	0.457 91	0.604 71	0.713 70	0.794 11	0.964 82	0.995 25	0.999 53	0.999 97	1	
20	20										0.000 00	20	20
	19									0.000 00	0.000 02	19	
	18									0.000 01	0.000 20	18	
	17								0.000 00	0.000 05	0.001 29	17	
	16								0.000 01	0.000 32	0.005 91	16	
	15								0.000 04	0.001 61	0.020 69	15	
	14							0.000 00	0.000 26	0.006 47	0.057 66	14	
	13							0.000 02	0.001 28	0.021 03	0.131 59	13	
	12							0.000 10	0.005 14	0.056 53	0.251 72	12	

Table 2. (Continued)

$n \backslash k$	π	0.01	0.02	0.04	0.06	0.08	0.1	0.2	0.3	0.4	0.5	$\pi \backslash k$	n
11							0.000 00	0.000 56	0.017 14	0.127 52	0.411 90	11	
10						0.000 00	0.000 01	0.002 59	0.047 96	0.244 60	0.588 10	10	
9					0.000 00	0.000 01	0.000 06	0.009 98	0.113 33	0.404 40	0.748 28	9	
8				0.000 00	0.000 01	0.000 09	0.000 42	0.032 14	0.227 73	0.584 11	0.868 41	8	
7				0.000 01	0.000 11	0.000 64	0.002 39	0.086 69	0.391 99	0.749 99	0.942 34	7	
6			0.000 00	0.000 10	0.000 87	0.003 80	0.011 25	0.195 79	0.583 63	0.874 40	0.979 31	6	
5		0.000 00	0.000 04	0.000 96	0.005 63	0.018 34	0.043 17	0.370 35	0.762 49	0.949 05	0.994 09	5	
4		0.000 04	0.000 60	0.007 41	0.028 97	0.070 62	0.132 95	0.588 55	0.892 91	0.984 04	0.998 71	4	
3		0.001 00	0.007 07	0.043 86	0.114 97	0.212 05	0.323 07	0.793 92	0.964 52	0.996 39	0.999 80	3	
2		0.016 86	0.059 90	0.189 66	0.339 55	0.483 14	0.608 25	0.930 82	0.992 36	0.999 48	0.999 98	2	
1		0.182 09	0.332 39	0.558 00	0.709 89	0.811 31	0.878 42	0.988 47	0.999 20	0.999 96	1.000 00	1	
25	25											25	25
	24										0.000 00		24
	23										0.000 01		23
	22									0.000 00	0.000 08		22
	21									0.000 01	0.000 46		21
	20									0.000 05	0.002 04		20
	19								0.000 00	0.000 28	0.007 32		19
	18								0.000 02	0.001 21	0.021 64		18
	17								0.000 10	0.004 33	0.053 88		17
	16							0.000 00	0.000 45	0.013 17	0.114 76		16
	15							0.000 01	0.001 78	0.034 39	0.212 18		15
	14							0.000 08	0.005 99	0.077 80	0.345 02		14
	13							0.000 37	0.017 47	0.153 77	0.500 00		13

Table 2. (Continued)

n	k	π	0.01	0.02	0.04	0.06	0.08	0.1	0.2	0.3	0.4	0.5	π	k	n
	12							0.000 00	0.001 54	0.044 25	0.267 72	0.654 98			12
	11						0.000 00	0.000 01	0.005 56	0.097 80	0.414 23	0.787 82			11
	10					0.000 00	0.000 01	0.000 08	0.017 33	0.189 44	0.575 38	0.885 24			10
	9					0.000 01	0.000 08	0.000 46	0.046 77	0.323 07	0.726 47	0.946 12			9
	8				0.000 00	0.000 07	0.000 52	0.002 26	0.109 12	0.488 15	0.846 45	0.978 36			8
	7			0.000 00	0.000 04	0.000 51	0.002 77	0.009 48	0.219 96	0.659 35	0.926 43	0.992 68			7
	6			0.000 01	0.000 38	0.003 06	0.012 29	0.033 40	0.383 31	0.806 51	0.970 64	0.997 96			6
	5	0.000 00	0.000 12	0.002 78	0.015 05	0.045 14	0.097 99	0.579 33	0.909 53	0.990 53	0.999 54				5
	4	0.000 11	0.001 45	0.016 52	0.059 76	0.135 09	0.236 41	0.766 01	0.966 76	0.997 63	0.999 92				4
	3	0.001 95	0.013 24	0.076 48	0.187 11	0.323 17	0.462 91	0.901 77	0.991 04	0.999 57	0.999 99				3
	2	0.025 76	0.088 65	0.264 19	0.447 34	0.605 28	0.728 79	0.972 61	0.998 43	0.999 95	1.000 00				2
	1	0.222 18	0.396 54	0.639 60	0.787 09	0.875 64	0.928 21	0.996 22	0.999 87	1.000 00	1.000 00				1
30	30														30
	29														29
	28														28
	27											0.000 00			27
	26											0.000 03			26
	25										0.000 00	0.000 16			25
	24										0.000 01	0.000 72			24
	23											0.000 05	0.002 61		23
	22									0.000 00	0.000 22	0.008 06			22
	21									0.000 01	0.000 86	0.021 39			21
	20									0.000 04	0.002 85	0.049 37			20
	19									0.000 16	0.008 30	0.100 24			19
	18								0.000 00	0.000 63	0.021 24	0.180 80			18

Table 2. (Continued)

$n \backslash k$	π	0.01	0.02	0.04	0.06	0.08	0.1	0.2	0.3	0.4	0.5	$\pi \backslash k$	n
17								0.000 01	0.002 12	0.048 11	0.292 33	17	
16								0.000 05	0.006 37	0.097 06	0.427 77	16	
15								0.000 23	0.016 94	0.175 37	0.572 23	15	
14								0.000 90	0.040 05	0.285 50	0.707 67	14	
13							0.000 00	0.003 11	0.084 47	0.421 53	0.819 20	13	
12						0.000 00	0.000 02	0.009 49	0.159 32	0.568 91	0.899 76	12	
11					0.000 00	0.000 01	0.000 09	0.025 62	0.269 63	0.708 53	0.950 63	11	
10					0.000 01	0.000 07	0.000 45	0.061 09	0.411 19	0.823 71	0.978 61	10	
9				0.000 00	0.000 05	0.000 41	0.002 02	0.128 65	0.568 48	0.905 99	0.991 94	9	
8				0.000 02	0.000 30	0.001 97	0.007 78	0.239 21	0.718 62	0.956 48	0.997 39	8	
7			0.000 00	0.000 15	0.001 67	0.008 25	0.025 83	0.393 03	0.840 48	0.982 82	0.999 28	7	
6	0.000 00	0.000 03	0.001 06	0.007 95	0.029 29	0.073 19	0.572 49	0.923 41	0.994 34	0.999 84		6	
5	0.000 01	0.000 30	0.006 32	0.031 54	0.087 36	0.175 49	0.744 77	0.969 85	0.998 49	0.999 97		5	
4	0.000 22	0.002 89	0.030 59	0.102 62	0.215 79	0.352 56	0.877 29	0.990 68	0.999 69	1.000 00		4	
3	0.003 32	0.021 72	0.116 90	0.267 60	0.434 60	0.588 60	0.955 82	0.997 89	0.999 95	1.000 00		3	
2	0.036 15	0.120 55	0.338 82	0.544 53	0.704 21	0.816 30	0.989 18	0.999 69	1.000 00	1.000 00		2	
1	0.260 30	0.454 52	0.706 14	0.843 74	0.918 03	0.957 61	0.998 76	1.000 00	1.000 00	1.000 00		1	

Table 3. The confidence interval of π in binomial distribution.

$1 - \alpha = 0.95$														
$k \backslash n - k$	1	2	3	4	5	6	7	8	9	10	12	14	16	$n - k \backslash k$
0	0.975	0.842	0.708	0.602	0.522	0.459	0.410	0.369	0.336	0.308	0.265	0.232	0.206	0
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
1	0.987	0.906	0.806	0.716	0.641	0.579	0.527	0.483	0.445	0.413	0.360	0.319	0.387	1
	0.013	0.008	0.006	0.005	0.004	0.004	0.003	0.003	0.003	0.002	0.002	0.002	0.001	
2	0.992	0.932	0.853	0.777	0.710	0.651	0.600	0.556	0.518	0.434	0.428	0.383	0.347	2
	0.094	0.068	0.053	0.043	0.037	0.032	0.028	0.025	0.023	0.021	0.018	0.016	0.014	
3	0.994	0.947	0.882	0.816	0.755	0.701	0.652	0.610	0.572	0.538	0.481	0.434	0.396	3
	0.194	0.147	0.118	0.099	0.085	0.075	0.067	0.060	0.055	0.050	0.043	0.038	0.034	
4	0.995	0.957	0.901	0.843	0.788	0.738	0.692	0.651	0.614	0.581	0.524	0.476	0.437	4
	0.284	0.223	0.184	0.157	0.137	0.122	0.109	0.099	0.091	0.084	0.073	0.064	0.057	
5	0.996	0.963	0.915	0.863	0.813	0.766	0.723	0.684	0.649	0.616	0.560	0.512	0.471	5
	0.359	0.290	0.245	0.212	0.187	0.167	0.151	0.139	0.128	0.118	0.103	0.091	0.082	
6	0.996	0.968	0.925	0.878	0.833	0.789	0.749	0.711	0.677	0.646	0.590	0.543	0.502	6
	0.421	0.349	0.299	0.262	0.234	0.211	0.192	0.177	0.163	0.152	0.133	0.119	0.107	
7	0.997	0.972	0.933	0.891	0.849	0.808	0.770	0.734	0.701	0.671	0.616	0.570	0.529	7
	0.473	0.400	0.348	0.308	0.277	0.251	0.230	0.213	0.198	0.184	0.163	0.146	0.132	

Table 3. (Continued)

$k \backslash n-k$	1	2	3	4	5	6	7	8	9	10	12	14	16	$n-k$
8	0.997	0.975	0.940	0.901	0.861	0.823	0.787	0.753	0.722	0.692	0.639	0.593	0.553	8
9	0.517	0.444	0.390	0.349	0.316	0.289	0.266	0.247	0.230	0.215	0.191	0.172	0.156	9
	0.997	0.977	0.945	0.909	0.872	0.837	0.802	0.770	0.740	0.711	0.660	0.615	0.575	
10	0.555	0.482	0.428	0.386	0.351	0.323	0.299	0.278	0.260	0.244	0.218	0.197	0.180	10
	0.998	0.979	0.950	0.916	0.882	0.848	0.816	0.785	0.756	0.728	0.678	0.634	0.595	
12	0.587	0.516	0.462	0.419	0.384	0.354	0.329	0.308	0.289	0.272	0.244	0.221	0.202	12
	0.998	0.982	0.957	0.927	0.897	0.867	0.837	0.809	0.782	0.756	0.709	0.666	0.628	
14	0.640	0.572	0.519	0.476	0.440	0.410	0.384	0.361	0.340	0.322	0.291	0.266	0.245	14
	0.998	0.984	0.962	0.936	0.909	0.881	0.854	0.828	0.803	0.779	0.734	0.694	0.657	
16	0.681	0.617	0.566	0.524	0.488	0.457	0.430	0.407	0.385	0.366	0.334	0.306	0.283	16
	0.999	0.986	0.966	0.943	0.918	0.893	0.868	0.844	0.820	0.798	0.755	0.717	0.681	
18	0.713	0.653	0.604	0.563	0.529	0.498	0.471	0.447	0.425	0.405	0.372	0.343	0.319	18
	0.999	0.988	0.970	0.948	0.925	0.902	0.879	0.857	0.835	0.814	0.773	0.736	0.702	
20	0.740	0.683	0.637	0.597	0.564	0.533	0.506	0.482	0.460	0.440	0.406	0.376	0.351	20
	0.999	0.989	0.972	0.953	0.932	0.910	0.889	0.868	0.847	0.827	0.789	0.753	0.720	
	0.762	0.708	0.664	0.626	0.593	0.564	0.537	0.513	0.492	0.472	0.437	0.407	0.381	

Table 3. (Continued)

$k \backslash n - k$	1	2	3	4	5	6	7	8	9	10	12	14	16	$n - k \backslash k$
22	0.999	0.990	0.975	0.956	0.937	0.917	0.897	0.877	0.858	0.839	0.803	0.768	0.737	22
	0.781	0.730	0.688	0.651	0.619	0.590	0.565	0.541	0.519	0.500	0.465	0.434	0.408	
24	0.999	0.991	0.976	0.960	0.942	0.923	0.904	0.885	0.867	0.849	0.814	0.782	0.751	24
	0.797	0.749	0.708	0.673	0.642	0.614	0.589	0.566	0.545	0.525	0.490	0.460	0.433	
26	0.999	0.991	0.978	0.962	0.945	0.928	0.910	0.893	0.875	0.858	0.825	0.794	0.764	26
	0.810	0.765	0.726	0.693	0.663	0.636	0.611	0.588	0.567	0.548	0.513	0.483	0.456	
28	0.999	0.992	0.980	0.965	0.949	0.932	0.916	0.899	0.882	0.866	0.834	0.804	0.776	28
	0.822	0.779	0.743	0.710	0.681	0.655	0.631	0.609	0.588	0.569	0.535	0.504	0.478	
30	0.999	0.992	0.981	0.967	0.952	0.936	0.920	0.904	0.889	0.873	0.843	0.814	0.786	30
	0.833	0.792	0.757	0.725	0.697	0.672	0.649	0.627	0.607	0.588	0.554	0.524	0.498	
40	0.999	0.994	0.985	0.975	0.963	0.951	0.938	0.925	0.912	0.900	0.875	0.850	0.827	40
	0.871	0.838	0.809	0.783	0.759	0.737	0.717	0.698	0.679	0.662	0.631	0.602	0.578	
60	1.000	0.996	0.990	0.983	0.975	0.966	0.957	0.948	0.939	0.929	0.911	0.893	0.874	60
	0.912	0.888	0.867	0.848	0.830	0.813	0.797	0.782	0.767	0.752	0.727	0.703	0.681	
100	1.000	0.998	0.994	0.989	0.984	0.979	0.973	0.967	0.962	0.955	0.943	0.931	0.919	100
	0.946	0.931	0.917	0.904	0.892	0.881	0.870	0.859	0.849	0.838	0.820	0.802	0.786	
200	1.000	0.999	0.997	0.995	0.992	0.989	0.986	0.983	0.980	0.977	0.970	0.964	0.957	200
	0.973	0.965	0.957	0.951	0.944	0.938	0.932	0.926	0.920	0.914	0.903	0.893	0.883	
500	1.000	1.000	0.999	0.998	0.997	0.996	0.995	0.993	0.992	0.991	0.988	0.985	0.982	500
	0.989	0.986	0.983	0.980	0.977	0.974	0.972	0.969	0.967	0.964	0.960	0.955	0.95	

Table 3. (Continued)

$1 - \alpha = 0.95$														
$n - k$ k	18	20	22	24	26	28	30	40	60	100	200	500	$n - k$ k	
0	0.185	0.168	0.154	0.142	0.132	0.123	0.116	0.088	0.060	0.036	0.018	0.007	0	
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1	
	0.260	0.238	0.219	0.203	0.190	0.178	0.167	0.129	0.088	0.054	0.027	0.011		
2	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000	2	
	0.317	0.292	0.270	0.251	0.235	0.221	0.208	0.162	0.112	0.069	0.035	0.014		
3	0.012	0.011	0.010	0.009	0.009	0.008	0.008	0.006	0.004	0.002	0.001	0.000	3	
	0.363	0.336	0.312	0.292	0.274	0.257	0.243	0.191	0.133	0.083	0.043	0.017		
4	0.030	0.028	0.025	0.024	0.022	0.020	0.019	0.015	0.010	0.006	0.003	0.001	4	
	0.403	0.374	0.349	0.327	0.307	0.290	0.275	0.217	0.152	0.096	0.049	0.020		
5	0.052	0.047	0.044	0.040	0.038	0.035	0.033	0.025	0.017	0.011	0.005	0.002	5	
	0.436	0.407	0.381	0.358	0.337	0.319	0.303	0.241	0.170	0.108	0.056	0.023		
6	0.075	0.068	0.063	0.058	0.055	0.051	0.048	0.037	0.025	0.016	0.008	0.003	6	
	0.467	0.436	0.410	0.386	0.364	0.345	0.328	0.263	0.187	0.119	0.062	0.026		
	0.098	0.090	0.083	0.077	0.072	0.068	0.064	0.049	0.034	0.021	0.011	0.004		

Table 3. (Continued)

$k \backslash n - k$	18	20	22	24	26	28	30	40	60	100	200	500	$n - k \backslash k$
7	0.494	0.463	0.435	0.411	0.389	0.369	0.351	0.283	0.203	0.130	0.068	0.028	7
	0.121	0.111	0.103	0.096	0.090	0.084	0.080	0.062	0.043	0.027	0.014	0.005	
8	0.518	0.487	0.459	0.434	0.412	0.391	0.373	0.302	0.218	0.141	0.074	0.031	8
	0.143	0.132	0.123	0.115	0.107	0.101	0.096	0.075	0.052	0.033	0.017	0.007	
9	0.540	0.508	0.481	0.455	0.433	0.412	0.393	0.321	0.233	0.151	0.080	0.033	9
	0.165	0.153	0.142	0.133	0.125	0.118	0.111	0.088	0.061	0.038	0.020	0.008	
10	0.560	0.528	0.500	0.475	0.452	0.431	0.412	0.338	0.248	0.162	0.086	0.036	10
	0.186	0.173	0.161	0.151	0.142	0.134	0.127	0.100	0.071	0.045	0.023	0.009	
12	0.594	0.563	0.535	0.510	0.487	0.465	0.446	0.369	0.273	0.180	0.097	0.040	12
	0.227	0.211	0.197	0.186	0.175	0.166	0.157	0.125	0.089	0.057	0.030	0.012	
14	0.624	0.593	0.566	0.540	0.517	0.496	0.476	0.398	0.297	0.198	0.107	0.045	14
	0.264	0.247	0.232	0.218	0.206	0.196	0.186	0.150	0.107	0.069	0.036	0.015	
16	0.649	0.619	0.592	0.567	0.544	0.522	0.502	0.422	0.319	0.214	0.117	0.050	16
	0.298	0.280	0.263	0.249	0.236	0.224	0.214	0.173	0.126	0.081	0.043	0.018	
18	0.671	0.642	0.615	0.590	0.568	0.547	0.527	0.445	0.340	0.230	0.127	0.054	18
	0.329	0.310	0.293	0.277	0.264	0.251	0.240	0.196	0.143	0.093	0.050	0.021	

Table 3. (Continued)

$k \backslash n-k$	18	20	22	24	26	28	30	40	60	100	200	500	$n-k \backslash k$
20	0.690	0.662	0.636	0.612	0.589	0.568	0.548	0.467	0.359	0.245	0.137	0.059	20
	0.358	0.338	0.320	0.304	0.289	0.276	0.264	0.217	0.160	0.105	0.057	0.024	
22	0.707	0.680	0.654	0.631	0.608	0.588	0.568	0.487	0.378	0.260	0.146	0.063	22
	0.385	0.364	0.346	0.329	0.314	0.300	0.287	0.237	0.177	0.117	0.063	0.027	
24	0.723	0.696	0.671	0.648	0.626	0.605	0.586	0.505	0.395	0.274	0.155	0.067	24
	0.410	0.388	0.369	0.352	0.337	0.322	0.309	0.257	0.193	0.128	0.070	0.030	
26	0.736	0.711	0.686	0.663	0.642	0.622	0.603	0.522	0.411	0.287	0.164	0.072	26
	0.432	0.411	0.392	0.374	0.358	0.343	0.330	0.276	0.208	0.140	0.077	0.033	
28	0.749	0.724	0.700	0.678	0.657	0.637	0.618	0.538	0.426	0.300	0.172	0.076	28
	0.453	0.432	0.412	0.395	0.378	0.363	0.349	0.294	0.223	0.153	0.083	0.036	
30	0.760	0.736	0.713	0.691	0.670	0.651	0.632	0.552	0.441	0.313	0.181	0.080	30
	0.473	0.452	0.432	0.414	0.397	0.382	0.368	0.311	0.237	0.162	0.090	0.039	
40	0.804	0.783	0.763	0.743	0.724	0.706	0.689	0.614	0.503	0.368	0.220	0.099	40
	0.555	0.533	0.513	0.495	0.478	0.462	0.448	0.386	0.303	0.213	0.122	0.053	
60	0.857	0.840	0.823	0.807	0.792	0.777	0.763	0.697	0.593	0.455	0.287	0.136	60
	0.660	0.641	0.622	0.605	0.589	0.574	0.559	0.497	0.407	0.300	0.181	0.083	
100	0.907	0.895	0.883	0.872	0.860	0.847	0.838	0.787	0.700	0.571	0.395	0.199	100
	0.770	0.755	0.740	0.726	0.713	0.700	0.687	0.632	0.545	0.429	0.280	0.138	
200	0.950	0.943	0.937	0.930	0.923	0.917	0.910	0.878	0.819	0.720	0.550	0.319	200
	0.873	0.863	0.854	0.845	0.836	0.828	0.819	0.780	0.713	0.605	0.450	0.253	
500	0.979	0.976	0.973	0.970	0.967	0.964	0.961	0.947	0.917	0.862	0.747	0.531	500
	0.946	0.941	0.937	0.933	0.928	0.924	0.920	0.901	0.864	0.801	0.681	0.469	

Table 3. (Continued)

$1 - \alpha = 0.99$														
$k \backslash n - k$	1	2	3	4	5	6	7	8	9	10	12	14	16	$n - k \backslash k$
0	0.995	0.929	0.829	0.734	0.653	0.586	0.531	0.484	0.445	0.411	0.357	0.315	0.282	0
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
1	0.997	0.959	0.889	0.815	0.746	0.685	0.632	0.585	0.544	0.509	0.449	0.402	0.363	1
	0.003	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000	
2	0.998	0.971	0.917	0.856	0.797	0.742	0.693	0.648	0.608	0.573	0.512	0.463	0.422	2
	0.041	0.029	0.023	0.019	0.016	0.014	0.012	0.011	0.010	0.009	0.008	0.007	0.006	
3	0.999	0.977	0.934	0.882	0.830	0.781	0.735	0.693	0.655	0.621	0.561	0.510	0.468	3
	0.111	0.083	0.066	0.055	0.047	0.042	0.037	0.033	0.030	0.028	0.024	0.021	0.019	
4	0.999	0.981	0.945	0.900	0.854	0.809	0.767	0.728	0.691	0.658	0.599	0.549	0.507	4
	0.185	0.144	0.118	0.100	0.087	0.077	0.069	0.062	0.057	0.053	0.045	0.040	0.036	
5	0.999	0.984	0.953	0.913	0.872	0.831	0.791	0.755	0.720	0.688	0.631	0.582	0.539	5
	0.254	0.203	0.170	0.146	0.128	0.114	0.103	0.094	0.087	0.080	0.070	0.062	0.055	
6	0.999	0.986	0.958	0.923	0.886	0.848	0.811	0.777	0.744	0.714	0.658	0.610	0.567	6
	0.315	0.258	0.219	0.191	0.169	0.152	0.138	0.127	0.117	0.109	0.095	0.085	0.076	
7	0.999	0.988	0.963	0.931	0.897	0.862	0.828	0.795	0.764	0.735	0.681	0.634	0.592	7
	0.368	0.307	0.265	0.233	0.209	0.189	0.172	0.159	0.147	0.137	0.121	0.108	0.097	

Table 3. (Continued)

$\begin{smallmatrix} n-k \\ k \end{smallmatrix}$	1	2	3	4	5	6	7	8	9	10	12	14	16	$n-k$ k
8	0.999	0.989	0.967	0.938	0.906	0.873	0.841	0.811	0.781	0.753	0.701	0.655	0.614	8
	0.415	0.352	0.307	0.272	0.245	0.223	0.205	0.189	0.176	0.165	0.146	0.131	0.119	
9	0.999	0.990	0.970	0.943	0.913	0.883	0.853	0.824	0.795	0.768	0.718	0.674	0.634	9
	0.456	0.392	0.345	0.309	0.280	0.256	0.236	0.219	0.205	0.192	0.171	0.154	0.140	
10	1.000	0.991	0.972	0.947	0.920	0.891	0.863	0.835	0.808	0.782	0.734	0.690	0.651	10
	0.491	0.427	0.379	0.342	0.312	0.286	0.265	0.247	0.232	0.218	0.195	0.176	0.161	
12	1.000	0.992	0.976	0.955	0.930	0.905	0.879	0.854	0.829	0.805	0.760	0.719	0.682	12
	0.551	0.488	0.439	0.401	0.369	0.342	0.319	0.299	0.282	0.266	0.240	0.218	0.200	
14	1.000	0.993	0.979	0.960	0.938	0.915	0.892	0.869	0.846	0.824	0.782	0.743	0.707	14
	0.598	0.537	0.490	0.451	0.418	0.380	0.366	0.345	0.326	0.310	0.281	0.257	0.237	
16	1.000	0.994	0.981	0.964	0.945	0.924	0.903	0.881	0.860	0.839	0.800	0.763	0.728	16
	0.637	0.578	0.532	0.493	0.461	0.433	0.408	0.386	0.366	0.349	0.318	0.293	0.272	
18	1.000	0.995	0.983	0.968	0.950	0.931	0.911	0.891	0.872	0.852	0.815	0.780	0.747	18
	0.669	0.613	0.568	0.530	0.498	0.469	0.445	0.422	0.402	0.384	0.353	0.326	0.304	
20	1.000	0.995	0.985	0.971	0.954	0.936	0.918	0.900	0.881	0.863	0.828	0.794	0.763	20
	0.696	0.642	0.599	0.562	0.530	0.502	0.478	0.455	0.435	0.417	0.384	0.357	0.334	

Table 3. (Continued)

$k \backslash n - k$	1	2	3	4	5	6	7	8	9	10	12	14	16	$n - k \backslash k$
22	1.000	0.996	0.986	0.973	0.958	0.941	0.924	0.907	0.890	0.873	0.839	0.807	0.777	22
	0.719	0.668	0.626	0.590	0.559	0.531	0.507	0.484	0.464	0.445	0.413	0.385	0.361	
24	1.000	0.996	0.987	0.975	0.961	0.946	0.930	0.913	0.897	0.881	0.849	0.819	0.789	24
	0.738	0.690	0.649	0.615	0.584	0.557	0.533	0.511	0.490	0.471	0.439	0.410	0.386	
26	1.000	0.996	0.988	0.977	0.963	0.949	0.934	0.919	0.903	0.888	0.858	0.829	0.800	26
	0.755	0.709	0.670	0.637	0.607	0.580	0.557	0.535	0.515	0.496	0.463	0.434	0.410	
28	1.000	0.996	0.989	0.978	0.966	0.952	0.938	0.924	0.909	0.894	0.866	0.838	0.811	28
	0.770	0.726	0.689	0.656	0.627	0.602	0.578	0.557	0.537	0.518	0.485	0.457	0.432	
30	1.000	0.997	0.989	0.980	0.968	0.955	0.942	0.928	0.914	0.900	0.873	0.846	0.820	30
	0.784	0.741	0.705	0.674	0.646	0.621	0.598	0.577	0.557	0.539	0.506	0.478	0.452	
40	1.000	0.998	0.992	0.984	0.975	0.965	0.955	0.944	0.933	0.921	0.899	0.876	0.854	40
	0.832	0.797	0.767	0.740	0.716	0.694	0.673	0.654	0.636	0.619	0.588	0.560	0.536	
60	1.000	0.998	0.995	0.989	0.983	0.976	0.969	0.961	0.953	0.945	0.928	0.912	0.895	60
	0.884	0.859	0.836	0.816	0.797	0.780	0.763	0.748	0.733	0.719	0.693	0.668	0.646	
100	1.000	0.999	0.997	0.993	0.990	0.985	0.981	0.976	0.971	0.965	0.955	0.943	0.932	100
	0.929	0.912	0.897	0.884	0.871	0.858	0.847	0.836	0.825	0.815	0.795	0.777	0.761	
200	1.000	0.999	0.998	0.997	0.995	0.992	0.990	0.988	0.985	0.982	0.976	0.940	0.964	200
	0.964	0.955	0.947	0.939	0.932	0.925	0.919	0.913	0.907	0.901	0.890	0.878	0.868	
500	1.000	1.000	0.999	0.999	0.998	0.997	0.996	0.995	0.994	0.993	0.990	0.988	0.985	500
	0.985	0.982	0.978	0.975	0.972	0.969	0.967	0.964	0.961	0.959	0.953	0.949	0.944	

Table 3. (Continued)

$1 - \alpha = 0.99$													
$k \backslash n - k$	18	20	22	24	26	28	30	40	60	100	200	500	$n - k \backslash k$
0	0.255	0.233	0.214	0.193	0.184	0.172	0.162	0.124	0.085	0.052	0.026	0.011	0
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
1	0.331	0.304	0.281	0.262	0.245	0.230	0.216	0.163	0.116	0.071	0.036	0.015	1
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
2	0.387	0.358	0.332	0.310	0.291	0.274	0.259	0.203	0.141	0.088	0.045	0.018	2
	0.005	0.005	0.004	0.004	0.004	0.004	0.003	0.002	0.002	0.001	0.001	0.000	
3	0.432	0.401	0.374	0.351	0.330	0.311	0.295	0.233	0.164	0.103	0.053	0.022	3
	0.017	0.015	0.014	0.013	0.012	0.011	0.011	0.008	0.005	0.003	0.002	0.001	
4	0.470	0.438	0.410	0.385	0.363	0.344	0.326	0.260	0.184	0.116	0.061	0.025	4
	0.032	0.029	0.027	0.025	0.023	0.022	0.020	0.016	0.011	0.007	0.003	0.001	
5	0.502	0.470	0.441	0.416	0.393	0.373	0.354	0.284	0.203	0.129	0.068	0.028	5
	0.050	0.046	0.042	0.039	0.037	0.034	0.032	0.025	0.017	0.010	0.005	0.002	
6	0.531	0.498	0.469	0.443	0.420	0.393	0.379	0.306	0.220	0.142	0.075	0.031	6
	0.069	0.064	0.059	0.054	0.051	0.048	0.045	0.035	0.024	0.015	0.008	0.003	
7	0.555	0.522	0.498	0.467	0.443	0.422	0.402	0.327	0.237	0.153	0.081	0.033	7
	0.089	0.082	0.076	0.070	0.066	0.062	0.058	0.045	0.031	0.019	0.010	0.004	

Table 3. (Continued)

$k \backslash n - k$	18	20	22	24	26	28	30	40	60	100	200	500	$n - k \backslash k$
8	0.578	0.545	0.516	0.489	0.465	0.443	0.423	0.346	0.252	0.164	0.087	0.036	8
	0.109	0.100	0.093	0.087	0.081	0.076	0.072	0.056	0.039	0.024	0.012	0.005	
9	0.598	0.565	0.536	0.510	0.485	0.463	0.443	0.364	0.267	0.175	0.093	0.039	9
	0.128	0.119	0.110	0.103	0.097	0.091	0.086	0.067	0.047	0.029	0.015	0.006	
10	0.616	0.583	0.555	0.529	0.504	0.482	0.461	0.381	0.281	0.185	0.099	0.041	10
	0.148	0.137	0.127	0.119	0.112	0.106	0.100	0.079	0.055	0.035	0.018	0.007	
12	0.647	0.616	0.587	0.561	0.537	0.515	0.494	0.412	0.307	0.205	0.110	0.047	12
	0.185	0.172	0.161	0.151	0.142	0.134	0.127	0.101	0.072	0.045	0.024	0.010	
14	0.674	0.643	0.615	0.590	0.566	0.543	0.522	0.440	0.332	0.223	0.122	0.051	14
	0.220	0.206	0.193	0.181	0.171	0.162	0.154	0.124	0.088	0.057	0.030	0.012	
16	0.696	0.666	0.639	0.614	0.590	0.568	0.548	0.464	0.354	0.239	0.132	0.056	16
	0.253	0.237	0.223	0.211	0.200	0.189	0.180	0.146	0.105	0.068	0.036	0.015	
18	0.716	0.687	0.661	0.636	0.612	0.591	0.570	0.486	0.374	0.255	0.142	0.061	18
	0.284	0.267	0.252	0.238	0.226	0.215	0.205	0.167	0.122	0.079	0.042	0.018	
20	0.733	0.705	0.679	0.655	0.632	0.611	0.591	0.507	0.394	0.271	0.152	0.066	20
	0.313	0.295	0.279	0.264	0.251	0.239	0.229	0.187	0.137	0.090	0.048	0.020	

Table 3. (Continued)

$k \backslash n-k$	18	20	22	24	26	28	30	40	60	100	200	500	$n-k \backslash k$
22	0.748	0.721	0.696	0.673	0.650	0.629	0.609	0.526	0.411	0.286	0.162	0.070	22
	0.339	0.321	0.304	0.289	0.274	0.263	0.251	0.207	0.153	0.101	0.054	0.023	
24	0.762	0.736	0.711	0.688	0.666	0.646	0.626	0.543	0.428	0.300	0.171	0.075	24
	0.364	0.345	0.327	0.312	0.298	0.285	0.273	0.226	0.168	0.112	0.061	0.026	
26	0.774	0.749	0.726	0.702	0.681	0.661	0.642	0.560	0.444	0.313	0.180	0.079	26
	0.388	0.368	0.350	0.334	0.319	0.306	0.293	0.244	0.183	0.122	0.067	0.029	
28	0.785	0.761	0.737	0.715	0.694	0.675	0.656	0.575	0.459	0.326	0.189	0.083	28
	0.409	0.389	0.371	0.354	0.339	0.325	0.312	0.262	0.193	0.133	0.073	0.031	
30	0.795	0.771	0.749	0.727	0.707	0.688	0.669	0.589	0.473	0.339	0.197	0.088	30
	0.430	0.409	0.391	0.374	0.358	0.344	0.331	0.278	0.212	0.143	0.079	0.034	
40	0.833	0.813	0.793	0.774	0.756	0.738	0.722	0.646	0.534	0.394	0.237	0.108	40
	0.514	0.493	0.474	0.457	0.440	0.425	0.411	0.354	0.276	0.193	0.110	0.048	
60	0.878	0.863	0.847	0.832	0.817	0.802	0.788	0.724	0.620	0.479	0.305	0.145	60
	0.625	0.606	0.589	0.572	0.556	0.541	0.527	0.466	0.380	0.278	0.167	0.076	
100	0.921	0.910	0.899	0.883	0.878	0.867	0.857	0.807	0.722	0.593	0.407	0.209	100
	0.745	0.729	0.714	0.700	0.687	0.674	0.661	0.606	0.521	0.407	0.265	0.129	
200	0.958	0.952	0.946	0.939	0.933	0.927	0.921	0.890	0.833	0.735	0.565	0.332	200
	0.858	0.848	0.838	0.829	0.820	0.811	0.803	0.763	0.695	0.593	0.435	0.243	
500	0.982	0.980	0.977	0.974	0.971	0.969	0.966	0.952	0.924	0.871	0.757	0.541	500
	0.939	0.934	0.930	0.925	0.921	0.917	0.912	0.892	0.855	0.791	0.668	0.459	

Table 4. The confidence interval of λ in Poisson distribution.

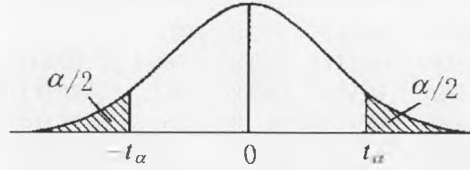
1 - α							1 - α						
N	90		95		99		N	90		95		99	
	U	L	U	L	U	L		U	L	U	L	U	L
0	0.00	3.00	0.00	3.69	0.00	5.30	15	9.25	23.10	8.40	24.74	6.89	28.16
1	0.05	4.74	0.03	5.57	0.01	7.43	16	10.04	24.30	9.15	25.98	7.57	29.48
2	0.36	6.30	0.24	7.22	0.10	9.27	17	10.83	25.50	9.90	27.22	8.25	30.79
3	0.82	7.75	0.62	8.77	0.34	10.98	18	11.63	26.69	10.67	28.45	8.94	32.09
4	1.37	9.15	1.09	10.24	0.67	12.59	19	12.44	27.88	11.44	29.67	9.64	33.38
5	1.97	10.51	1.62	11.67	1.08	14.15	20	13.25	29.06	12.22	30.89	10.35	34.67
6	2.61	11.84	2.20	13.06	1.54	15.66	21	14.07	30.24	13.00	32.10	11.07	35.95
7	3.29	13.15	2.81	14.42	2.04	17.13	22	14.89	31.41	13.79	33.31	11.79	37.22
8	3.98	14.43	3.45	15.76	2.57	18.58	23	15.72	32.59	14.58	34.51	12.52	38.48
9	4.70	15.71	4.12	17.08	3.13	20.00	24	16.55	33.75	15.38	35.71	13.26	39.74
10	5.43	16.96	4.80	18.39	3.72	21.40	25	17.38	34.92	16.18	36.90	14.00	41.00
11	6.17	18.21	5.49	19.68	4.32	22.78	26	18.22	36.08	16.98	38.10	14.74	42.25
12	6.92	19.44	6.20	20.96	4.94	24.14	27	19.06	37.23	17.79	39.28	15.49	43.50
13	7.69	20.67	6.92	22.23	5.58	25.50	28	19.90	38.39	18.61	40.47	16.25	44.74
14	8.46	21.89	7.65	23.49	6.23	26.84	29	20.75	39.54	19.42	41.65	17.00	45.98

Table 4. (Continued)

1 - α							1 - α						
90		95		99		N	90		95		99		
U	L	U	L	U	L		U	L	U	L	U	L	
70	56.83	85.40	54.57	88.44	50.33	94.58	85	70.42	101.80	67.89	105.10	63.13	111.76
71	57.73	86.50	55.45	89.56	51.17	95.73	86	71.34	102.89	68.79	106.21	63.99	112.90
72	58.63	87.60	56.34	90.67	52.02	96.88	87	72.25	103.98	69.68	107.31	64.85	114.04
73	59.54	88.69	57.22	91.79	52.87	98.03	88	73.16	105.06	70.58	108.42	65.72	115.17
74	60.44	89.79	58.11	92.90	53.72	99.18	89	74.07	106.15	71.47	109.52	66.58	116.31
75	61.35	90.89	58.99	94.01	54.57	100.33	90	74.98	107.24	72.37	110.63	67.44	117.45
76	62.25	91.98	59.88	95.13	55.42	101.48	91	75.90	108.32	73.27	111.73	68.31	118.58
77	63.16	93.07	60.77	96.24	56.28	102.62	92	76.81	109.41	74.16	112.83	69.17	119.71
78	64.06	94.17	61.66	97.35	57.13	103.77	93	77.73	110.50	75.06	113.93	70.04	120.85
79	64.97	95.26	62.55	98.46	57.98	104.91	94	78.64	111.58	75.96	115.03	70.91	121.98
80	65.88	96.35	63.44	99.57	58.84	106.06	95	79.56	112.66	76.86	116.13	71.77	123.11
81	66.79	97.44	64.33	100.68	59.70	107.20	96	80.47	113.75	77.76	117.23	72.64	124.24
82	67.70	98.53	65.22	101.78	60.55	108.34	97	81.39	114.83	78.66	118.33	73.51	125.37
83	68.60	99.62	66.11	102.89	61.41	109.48	98	82.30	115.91	79.56	119.43	74.38	126.50
84	69.51	100.71	67.00	104.00	62.27	110.62	99	83.22	117.00	80.46	120.53	75.25	127.63
							100	84.14	118.08	81.36	121.63	76.12	128.76

Table 5. Critical values of t distribution.

$$P(|t| > t_\alpha) = \alpha$$

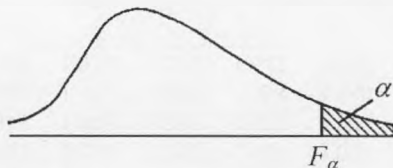


$\nu \backslash \alpha$	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001	$\alpha \backslash \nu$
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619	1
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598	2
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924	3
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610	4
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859	5
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959	6
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.993	3.499	5.405	7
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041	8
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781	9
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587	10
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437	11
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318	12
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221	13
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140	14
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073	15

Table 5. (Continued)

$\nu \backslash a$	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001	$a \backslash \nu$
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015	16
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965	17
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922	18
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883	19
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850	20
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819	21
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792	22
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767	23
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745	24
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725	25
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707	26
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690	27
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674	28
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659	29
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646	30
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551	40
60	0.126	0.254	0.387	0.527	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460	60
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373	120
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291	∞

Table 6.1. Critical values of the F distribution (upper tail).

$P(F > F_\alpha) = \alpha$


		$\alpha = 0.10$																			
		ν_1																			
ν_2		1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	200	500	∞		ν_2
1	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	61.2	61.7	62.3	62.7	63.0	63.2	63.3	63.3		1	
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.42	9.44	9.46	9.47	9.48	9.49	9.49	9.49		2	
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.20	5.18	5.17	5.15	5.14	5.14	5.14	5.13		3	
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.87	3.84	3.82	3.80	3.78	3.77	3.76	3.76		4	
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.24	3.21	3.17	3.15	3.13	3.12	3.11	3.11		5	
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.87	2.84	2.80	2.77	2.75	2.73	2.73	2.72		6	
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.63	2.59	2.56	2.52	2.50	2.48	2.48	2.47		7	
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.46	2.42	2.38	2.35	2.32	2.31	2.30	2.29		8	
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.34	2.30	2.25	2.22	2.19	2.17	2.17	2.16		9	
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.24	2.20	2.16	2.12	2.09	2.07	2.06	2.06		10	
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.17	2.12	2.08	2.04	2.01	1.99	1.98	1.97		11	
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.10	2.06	2.01	1.97	1.94	1.92	1.91	1.90		12	
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.05	2.01	1.96	1.92	1.88	1.86	1.85	1.85		13	
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.01	1.96	1.91	1.87	1.83	1.82	1.80	1.80		14	
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.97	1.92	1.87	1.83	1.79	1.77	1.76	1.76		15	

Table 6.1. (Continued)

$\alpha = 0.10$																			
ν_1																			
ν_2	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	200	500	∞	ν_2
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.94	1.89	1.84	1.79	1.76	1.74	1.73	1.72	16
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.91	1.86	1.81	1.76	1.73	1.71	1.69	1.69	17
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.89	1.84	1.78	1.74	1.70	1.68	1.67	1.66	18
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.86	1.81	1.76	1.71	1.67	1.65	1.64	1.63	19
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.84	1.79	1.74	1.69	1.65	1.63	1.62	1.61	20
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.81	1.76	1.70	1.65	1.61	1.59	1.58	1.57	22
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.78	1.73	1.67	1.62	1.58	1.56	1.54	1.53	24
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.76	1.71	1.65	1.59	1.55	1.53	1.51	1.50	26
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.74	1.69	1.63	1.57	1.53	1.50	1.49	1.48	28
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.72	1.67	1.61	1.55	1.51	1.48	1.47	1.46	30
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.66	1.61	1.54	1.48	1.43	1.41	1.39	1.38	40
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.63	1.57	1.50	1.44	1.39	1.36	1.34	1.33	50
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.60	1.54	1.48	1.41	1.36	1.33	1.31	1.29	60
80	2.77	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.71	1.68	1.57	1.51	1.44	1.38	1.32	1.28	1.26	1.24	80
100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.70	1.66	1.56	1.49	1.42	1.35	1.29	1.26	1.23	1.21	100
200	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.69	1.66	1.63	1.52	1.46	1.38	1.31	1.24	1.20	1.17	1.14	200
500	2.72	2.31	2.09	1.96	1.86	1.79	1.73	1.68	1.64	1.61	1.50	1.44	1.36	1.28	1.21	1.16	1.12	1.09	500
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.49	1.42	1.34	1.26	1.18	1.13	1.08	1.00	∞

Table 6.1. (Continued)

		$\alpha = 0.05$																			
		ν_1																			
ν_2		1	2	3	4	5	6	7	8	9	10	12	14	16	18	20					ν_2
1	161	200	216	225	230	234	237	239	241	242	244	244	245	246	247	248					1
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4					2
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.74	8.71	8.69	8.67	8.66					3
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.91	5.87	5.84	5.82	5.80					4
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.68	4.64	4.60	4.58	4.56					5
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	4.00	3.96	3.92	3.90	3.87					6
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.57	3.53	3.49	3.47	3.44					7
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.28	3.24	3.20	3.17	3.15					8
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.07	3.03	2.99	2.96	2.94					9
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.91	2.86	2.83	2.80	2.77					10
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.79	2.74	2.70	2.67	2.65					11
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.69	2.64	2.60	2.57	2.54					12
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.60	2.55	2.51	2.48	2.46					13
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.53	2.48	2.44	2.41	2.39					14
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.48	2.42	2.38	2.35	2.33					15
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.42	2.37	2.33	2.30	2.28					16
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.38	2.33	2.29	2.26	2.23					17
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.34	2.29	2.25	2.22	2.19					18
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.31	2.26	2.21	2.18	2.16					19
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.28	2.22	2.18	2.15	2.12					20

Table 6.1. (Continued)

$\alpha = 0.05$																
ν_1																
ν_2	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	ν_2
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.20	2.16	2.12	2.10	21
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.17	2.13	2.10	2.07	22
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.15	2.11	2.08	2.05	23
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.13	2.09	2.05	2.03	24
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.11	2.07	2.04	2.01	25
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.09	2.05	2.02	1.99	26
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.08	2.04	2.00	1.97	27
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.06	2.02	1.99	1.96	28
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.05	2.01	1.97	1.94	29
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.04	1.99	1.96	1.93	30
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	2.07	2.01	1.97	1.94	1.91	32
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	2.05	1.99	1.95	1.92	1.89	34
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11	2.03	1.98	1.93	1.90	1.87	36
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.02	1.96	1.92	1.88	1.85	38
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.95	1.90	1.87	1.84	40
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	1.99	1.94	1.89	1.86	1.83	42
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	1.98	1.92	1.88	1.84	1.81	44
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.15	2.09	2.04	1.97	1.91	1.87	1.83	1.80	46
48	4.04	3.19	2.80	2.57	2.41	2.29	2.21	2.14	2.08	2.03	1.96	1.90	1.86	1.82	1.79	48
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.89	1.85	1.81	1.78	50

Table 6.1. (Continued)

		$\alpha = 0.05$															
		ν_1															
ν_2		1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	∞
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.86	1.82	1.77	1.73	1.70	60
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.82	1.77	1.73	1.71	1.68	80
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.79	1.75	1.73	1.71	1.68	100
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.96	1.91	1.83	1.77	1.73	1.71	1.69	1.66	125
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.82	1.76	1.71	1.71	1.67	1.64	150
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.80	1.74	1.69	1.69	1.66	1.62	200
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.78	1.72	1.68	1.68	1.64	1.61	300
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.77	1.71	1.66	1.66	1.62	1.59	500
1000	3.85	3.00	2.60	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.70	1.65	1.65	1.61	1.58	1000
∞	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.69	1.64	1.64	1.60	1.57	∞

Table 6.1. (Continued)

$\alpha = 0.05$															
ν_2	ν_1														
	22	24	26	28	30	35	40	45	50	60	80	100	200	500	∞
1	249	249	249	250	250	251	251	251	252	252	253	253	254	254	254
2	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5
3	8.65	8.64	8.63	8.62	8.62	8.60	8.59	8.59	8.58	8.57	8.56	8.55	8.54	8.53	8.53
4	5.79	5.77	5.76	5.75	5.75	5.73	5.72	5.71	5.70	5.69	5.67	5.66	5.65	5.64	5.63
5	4.54	4.53	4.52	4.50	4.50	4.48	4.46	4.45	4.44	4.43	4.41	4.41	4.39	4.37	4.37
6	3.86	3.84	3.83	3.82	3.81	3.79	3.77	3.76	3.75	3.74	3.72	3.71	3.69	3.68	3.67
7	3.43	3.41	3.40	3.39	3.38	3.36	3.34	3.33	3.32	3.30	3.29	3.27	3.25	3.24	3.23
8	3.13	3.12	3.10	3.09	3.08	3.06	3.04	3.03	3.02	3.01	2.99	2.97	2.95	2.94	2.93
9	2.92	2.90	2.89	2.87	2.86	2.84	2.83	2.81	2.80	2.79	2.77	2.76	2.73	2.72	2.71
10	2.75	2.74	2.72	2.71	2.70	2.68	2.66	2.65	2.64	2.62	2.60	2.59	2.56	2.55	2.54
11	2.63	2.61	2.59	2.58	2.57	2.55	2.53	2.52	2.51	2.49	2.47	2.46	2.43	2.42	2.40
12	2.52	2.51	2.49	2.48	2.47	2.44	2.43	2.41	2.40	2.38	2.36	2.35	2.32	2.31	2.30
13	2.44	2.42	2.41	2.39	2.38	2.36	2.34	2.33	2.31	2.30	2.27	2.26	2.23	2.22	2.21
14	2.37	2.35	2.33	2.32	2.31	2.28	2.27	2.25	2.24	2.22	2.20	2.19	2.16	2.14	2.13
15	2.31	2.29	2.27	2.36	2.25	2.22	2.20	2.19	2.18	2.16	2.14	2.12	2.10	2.08	2.07
16	2.25	2.24	2.22	2.21	2.19	2.17	2.15	2.14	2.12	2.11	2.08	2.07	2.04	2.02	2.01
17	2.21	2.19	2.17	2.16	2.15	2.12	2.10	2.09	2.08	2.06	2.03	2.02	1.99	1.97	1.96
18	2.17	2.15	2.13	2.12	2.11	2.08	2.06	2.05	2.04	2.02	1.99	1.98	1.95	1.93	1.92
19	2.13	2.11	2.10	2.08	2.07	2.05	2.03	2.01	2.00	1.98	1.96	1.94	1.91	1.89	1.88
20	2.10	2.08	2.07	2.05	2.04	2.01	1.99	1.98	1.97	1.95	1.92	1.91	1.88	1.86	1.84

Table 6.1. (Continued)

		$\alpha = 0.05$														
		ν_1														
ν_2		22	24	26	28	30	35	40	45	50	60	80	100	200	500	∞
21	2.07	2.05	2.04	2.02	2.01	1.98	1.96	1.95	1.94	1.92	1.89	1.88	1.84	1.83	1.81	21
22	2.05	2.03	2.01	2.00	1.98	1.96	1.94	1.92	1.91	1.89	1.86	1.85	1.82	1.80	1.78	22
23	2.02	2.01	1.99	1.97	1.96	1.93	1.91	1.90	1.88	1.86	1.84	1.82	1.79	1.77	1.76	23
24	2.00	1.98	1.97	1.95	1.94	1.91	1.89	1.88	1.86	1.84	1.82	1.80	1.77	1.75	1.73	24
25	1.98	1.96	1.95	1.93	1.92	1.89	1.87	1.86	1.84	1.82	1.80	1.78	1.75	1.73	1.71	25
26	1.97	1.95	1.93	1.91	1.90	1.87	1.85	1.84	1.82	1.80	1.78	1.76	1.73	1.71	1.69	26
27	1.95	1.93	1.91	1.90	1.88	1.86	1.84	1.82	1.81	1.79	1.76	1.74	1.71	1.69	1.67	27
28	1.93	1.91	1.90	1.88	1.87	1.84	1.82	1.80	1.79	1.77	1.74	1.73	1.69	1.67	1.65	28
29	1.92	1.90	1.88	1.87	1.85	1.83	1.81	1.79	1.77	1.75	1.73	1.71	1.67	1.65	1.64	29
30	1.91	1.89	1.87	1.85	1.84	1.81	1.79	1.77	1.76	1.74	1.71	1.70	1.66	1.64	1.62	30
32	1.88	1.86	1.85	1.83	1.82	1.79	1.77	1.75	1.74	1.71	1.69	1.67	1.63	1.61	1.59	32
34	1.86	1.84	1.82	1.81	1.80	1.77	1.75	1.73	1.71	1.69	1.66	1.65	1.61	1.59	1.57	34
36	1.85	1.82	1.81	1.79	1.78	1.75	1.73	1.71	1.69	1.67	1.64	1.62	1.59	1.56	1.55	36
38	1.83	1.81	1.79	1.77	1.76	1.73	1.71	1.69	1.68	1.65	1.62	1.61	1.57	1.54	1.53	38
40	1.81	1.79	1.77	1.76	1.74	1.72	1.69	1.67	1.66	1.64	1.61	1.59	1.55	1.53	1.51	40

Table 6.1. (Continued)

$\alpha = 0.05$																
ν_1																
ν_2	22	24	26	28	30	35	40	45	50	60	80	100	200	500	∞	ν_2
42	1.80	1.78	1.76	1.75	1.73	1.70	1.68	1.66	1.65	1.62	1.59	1.57	1.53	1.51	1.49	42
44	1.79	1.77	1.75	1.73	1.72	1.69	1.67	1.65	1.63	1.61	1.58	1.56	1.52	1.49	1.48	44
46	1.78	1.76	1.74	1.72	1.71	1.68	1.65	1.64	1.62	1.60	1.57	1.55	1.51	1.48	1.46	46
48	1.77	1.75	1.73	1.71	1.70	1.67	1.64	1.62	1.61	1.59	1.56	1.54	1.49	1.47	1.45	48
50	1.76	1.74	1.72	1.70	1.69	1.66	1.63	1.61	1.60	1.58	1.54	1.52	1.48	1.46	1.44	50
60	1.72	1.70	1.68	1.66	1.65	1.62	1.59	1.57	1.56	1.53	1.50	1.48	1.44	1.41	1.39	60
80	1.68	1.65	1.63	1.62	1.60	1.57	1.54	1.52	1.51	1.48	1.45	1.43	1.38	1.35	1.32	80
100	1.65	1.63	1.61	1.59	1.57	1.54	1.52	1.49	1.48	1.45	1.41	1.39	1.34	1.31	1.28	100
125	1.63	1.60	1.58	1.57	1.55	1.52	1.49	1.47	1.45	1.42	1.39	1.36	1.31	1.27	1.25	125
150	1.61	1.59	1.57	1.55	1.54	1.50	1.48	1.45	1.44	1.41	1.37	1.34	1.29	1.25	1.22	150
200	1.60	1.57	1.55	1.53	1.52	1.48	1.46	1.43	1.41	1.39	1.35	1.32	1.26	1.22	1.19	200
300	1.58	1.55	1.53	1.51	1.50	1.46	1.43	1.41	1.39	1.36	1.32	1.30	1.23	1.19	1.15	300
500	1.56	1.54	1.52	1.50	1.48	1.45	1.42	1.40	1.38	1.35	1.30	1.28	1.21	1.16	1.11	500
1000	1.55	1.53	1.51	1.49	1.47	1.43	1.41	1.38	1.36	1.33	1.29	1.26	1.19	1.13	1.08	1000
∞	1.54	1.52	1.50	1.48	1.46	1.42	1.39	1.37	1.35	1.32	1.27	1.24	1.17	1.11	1.00	∞

Table 6.1. (Continued)

$\alpha = 0.01$																
$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	ν_2
1	4 052	5 000	5 403	5 625	5 763	5 859	5 928	5 981	6 022	6 056	6 106	6 143	6 170	6 192	6 209	1
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	2
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.8	26.8	26.7	3
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.2	14.1	14.0	4
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.77	9.68	9.61	9.55	5
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.60	7.52	7.45	7.40	6
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.36	6.28	6.21	6.16	7
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.56	5.48	5.41	5.36	8
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	5.01	4.92	4.86	4.81	9
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.60	4.52	4.46	4.41	10
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.29	4.21	4.15	4.10	11
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.05	3.97	3.91	3.86	12
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.86	3.78	3.72	3.66	13
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.70	3.62	3.56	3.51	14
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.56	3.49	3.42	3.37	15

Table 6.1. (Continued)

$\alpha = 0.01$																
ν_2	ν_1															ν_2
	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.45	3.37	3.31	3.26	16
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.35	3.27	3.21	3.16	17
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.27	3.19	3.13	3.08	18
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.19	3.12	3.05	3.00	19
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.13	3.05	2.99	2.94	20
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.07	2.99	2.93	2.88	21
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	3.02	2.94	2.88	2.83	22
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.97	2.89	2.83	2.78	23
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.93	2.85	2.79	2.74	24
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.89	2.81	2.75	2.70	25
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.86	2.78	2.72	2.66	26
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.82	2.75	2.68	2.63	27
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.79	2.72	2.65	2.60	28
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.77	2.69	2.63	2.57	29
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.74	2.66	2.60	2.55	30

Table 6.1. (Continued)

$\alpha = 0.01$																			
v_1																			
v_2	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	v_2			
32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93	2.80	2.70	2.62	2.55	2.50	32			
34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98	2.89	2.76	2.66	2.58	2.51	2.46	34			
36	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95	2.86	2.72	2.62	2.54	2.48	2.43	36			
38	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92	2.83	2.69	2.59	2.51	2.45	2.40	38			
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.56	2.48	2.42	2.37	40			
42	7.28	5.15	4.29	3.80	3.49	3.27	3.10	2.97	2.86	2.78	2.64	2.54	2.46	2.40	2.34	42			
44	7.25	5.12	4.26	3.78	3.47	3.24	3.08	2.95	2.84	2.75	2.62	2.52	2.44	2.37	2.32	44			
46	7.22	5.10	4.24	3.76	3.44	3.22	3.06	2.93	2.82	2.73	2.60	2.50	2.42	2.35	2.30	46			
48	7.19	5.08	4.22	3.74	3.43	3.20	3.04	2.91	2.80	2.71	2.58	2.48	2.40	2.33	2.28	48			
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.46	2.38	2.32	2.27	50			
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.39	2.31	2.25	2.20	60			
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.42	2.31	2.23	2.17	2.12	80			
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.37	2.26	2.19	2.12	2.07	100			
125	6.84	4.78	3.94	3.47	3.17	2.95	2.79	2.66	2.55	2.47	2.33	2.23	2.15	2.08	2.03	125			
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.31	2.20	2.12	2.06	2.00	150			
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.27	2.17	2.09	2.03	1.97	200			
300	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.24	2.14	2.06	1.99	1.94	300			
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	2.22	2.12	2.04	1.97	1.92	500			
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.20	2.10	2.02	1.95	1.90	1000			
∞	6.64	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.08	2.00	1.93	1.88	∞			

Table 6.1. (Continued)

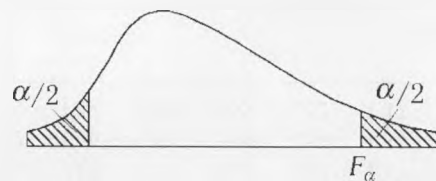
$\alpha = 0.01$																
ν_1																
ν_2	22	24	26	28	30	35	40	45	50	60	80	100	200	500	∞	ν_2
1	6 223	6 235	6 245	6 253	6 261	6 276	6 287	6 296	6 303	6 313	6 326	6 334	6 350	6 360	6 366	1
2	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	2
3	26.6	26.6	26.6	26.5	26.5	26.5	26.4	26.4	26.4	26.3	26.3	26.2	26.2	26.1	26.1	3
4	14.0	13.9	13.9	13.9	13.8	13.8	13.7	13.7	13.7	13.7	13.6	13.6	13.5	13.5	13.5	4
5	9.51	9.47	9.43	9.40	9.38	9.33	9.29	9.26	9.24	9.20	9.16	9.13	9.08	9.04	9.02	5
6	7.35	7.31	7.28	7.25	7.23	7.18	7.14	7.11	7.09	7.06	7.01	6.99	6.93	6.90	6.88	6
7	6.11	6.07	6.04	6.02	5.99	5.94	5.91	5.88	5.86	5.82	5.78	5.75	5.70	5.67	5.65	7
8	5.32	5.28	5.25	5.22	5.20	5.15	5.12	5.09	5.07	5.03	4.99	4.96	4.91	4.88	4.86	8
9	4.77	4.73	4.70	4.67	4.65	4.60	4.57	4.54	4.52	4.48	4.44	4.41	4.36	4.33	4.31	9
10	4.36	4.33	4.30	4.27	4.25	4.20	4.17	4.14	4.12	4.08	4.04	4.01	3.96	3.93	3.91	10
11	4.06	4.02	3.99	3.96	3.94	3.89	3.86	3.83	3.81	3.78	3.73	3.71	3.66	3.62	3.60	11
12	3.82	3.78	3.75	3.72	3.70	3.65	3.62	3.59	3.57	3.54	3.49	3.47	3.41	3.38	3.36	12
13	3.62	3.59	3.56	3.53	3.51	3.46	3.43	3.40	3.38	3.34	3.30	3.27	3.22	3.19	3.17	13
14	3.46	3.43	3.40	3.37	3.35	3.30	3.27	3.24	3.22	3.18	3.14	3.11	3.06	3.03	3.00	14
15	3.33	3.29	3.26	3.24	3.21	3.17	3.13	3.10	3.08	3.05	3.00	2.98	2.92	2.89	2.87	15
16	3.22	3.18	3.15	3.12	3.10	3.05	3.02	2.99	2.97	2.93	2.89	2.86	2.81	2.78	2.75	16
17	3.12	3.08	3.05	3.03	3.00	2.96	2.92	2.89	2.87	2.83	2.79	2.76	2.71	2.68	2.65	17
18	3.03	3.00	2.97	2.94	2.92	2.87	2.84	2.81	2.78	2.75	2.70	2.68	2.62	2.59	2.57	18
19	2.96	2.92	2.89	2.87	2.84	2.80	2.76	2.73	2.71	2.67	2.63	2.60	2.55	2.51	2.49	19
20	2.90	2.86	2.83	2.80	2.78	2.73	2.69	2.67	2.64	2.61	2.56	2.54	2.48	2.44	2.42	20

Table 6.1. (Continued)

$\alpha = 0.01$																
		ν_1														
ν_2		22	24	26	28	30	35	40	45	50	60	80	100	200	500	∞
21	2.84	2.80	2.77	2.74	2.72	2.67	2.64	2.61	2.58	2.55	2.50	2.48	2.42	2.38	2.36	2.36
22	2.78	2.75	2.72	2.69	2.67	2.62	2.58	2.55	2.53	2.50	2.45	2.42	2.36	2.33	2.31	2.32
23	2.74	2.70	2.67	2.64	2.62	2.57	2.54	2.51	2.48	2.45	2.40	2.37	2.32	2.28	2.26	2.23
24	2.70	2.66	2.63	2.60	2.58	2.53	2.49	2.46	2.44	2.40	2.36	2.33	2.27	2.24	2.21	2.24
25	2.66	2.62	2.59	2.56	2.54	2.49	2.45	2.42	2.40	2.36	2.32	2.29	2.23	2.19	2.17	2.25
26	2.62	2.58	2.55	2.53	2.50	2.45	2.42	2.39	2.36	2.33	2.28	2.25	2.19	2.16	2.13	2.26
27	2.59	2.55	2.52	2.49	2.47	2.42	2.38	2.35	2.33	2.29	2.25	2.22	2.16	2.12	2.10	2.27
28	2.56	2.52	2.49	2.46	2.44	2.39	2.35	2.32	2.30	2.26	2.22	2.19	2.13	2.09	2.06	2.28
29	2.53	2.49	2.46	2.44	2.41	2.36	2.33	2.30	2.27	2.23	2.19	2.16	2.10	2.06	2.03	2.29
30	2.51	2.47	2.44	2.41	2.39	2.34	2.30	2.27	2.25	2.21	2.16	2.13	2.07	2.03	2.01	3.0
32	2.46	2.42	2.39	2.36	2.34	2.29	2.25	2.22	2.20	2.16	2.11	2.08	2.02	1.98	1.96	3.2
34	2.42	2.38	2.35	2.32	2.30	2.25	2.21	2.18	2.16	2.12	2.07	2.04	1.98	1.94	1.91	3.4
36	2.38	2.35	2.32	2.29	2.26	2.21	2.18	2.14	2.12	2.08	2.03	2.00	1.94	1.90	1.87	3.6
38	2.35	2.32	2.28	2.26	2.23	2.18	2.14	2.11	2.09	2.05	2.00	1.97	1.90	1.86	1.84	3.8
40	2.33	2.29	2.26	2.23	2.20	2.15	2.11	2.08	2.06	2.02	1.97	1.94	1.87	1.83	1.80	4.0

Table 6.1. (Continued)

$\alpha = 0.01$																
ν_2	ν_1															ν_2
	22	24	26	28	30	35	40	45	50	60	80	100	200	500	∞	
42	2.30	2.26	2.23	2.20	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.91	1.85	1.80	1.78	42
44	2.28	2.24	2.21	2.18	2.15	2.10	2.07	2.03	2.01	1.97	1.92	1.89	1.82	1.78	1.75	44
46	2.26	2.22	2.19	2.16	2.13	2.08	2.04	2.01	1.99	1.95	1.90	1.86	1.80	1.76	1.73	46
48	2.24	2.20	2.17	2.14	2.12	2.06	2.02	1.99	1.97	1.93	1.88	1.84	1.78	1.73	1.70	48
50	2.22	2.18	2.15	2.12	2.10	2.05	2.01	1.97	1.95	1.91	1.86	1.82	1.76	1.71	1.68	50
60	2.15	2.12	2.08	2.05	2.03	1.98	1.94	1.90	1.88	1.84	1.78	1.75	1.68	1.63	1.60	60
80	2.07	2.03	2.00	1.97	1.94	1.89	1.85	1.82	1.79	1.75	1.69	1.65	1.58	1.53	1.49	80
100	2.02	1.98	1.94	1.92	1.89	1.84	1.80	1.76	1.74	1.69	1.63	1.60	1.52	1.47	1.43	100
125	1.98	1.94	1.91	1.88	1.85	1.80	1.76	1.72	1.69	1.65	1.59	1.55	1.47	1.41	1.37	125
150	1.96	1.92	1.88	1.85	1.83	1.77	1.73	1.69	1.66	1.62	1.56	1.52	1.43	1.38	1.33	150
200	1.93	1.89	1.85	1.82	1.79	1.74	1.69	1.66	1.63	1.58	1.52	1.48	1.39	1.33	1.28	200
300	1.89	1.85	1.82	1.79	1.76	1.70	1.66	1.62	1.59	1.55	1.48	1.44	1.35	1.28	1.22	300
500	1.87	1.83	1.79	1.76	1.74	1.68	1.63	1.60	1.57	1.52	1.45	1.41	1.31	1.23	1.16	500
1000	1.85	1.81	1.77	1.74	1.72	1.66	1.61	1.58	1.54	1.50	1.43	1.38	1.28	1.19	1.11	1000
∞	1.83	1.79	1.76	1.72	1.70	1.64	1.59	1.55	1.52	1.47	1.40	1.36	1.25	1.15	1.00	∞

Table 6.2. Critical values of F distribution (two tailed).

$$\alpha = 0.05$$

ν_1																			
ν_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	973.0	976.7	979.8	982.5	984.9	986.9	988.7	990.4	991.8
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.41	39.42	39.43	39.43	39.44	39.44	39.44	39.45
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.37	14.34	14.30	14.28	14.25	14.23	14.21	14.20	14.18
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.79	8.75	8.71	8.68	8.66	8.63	8.61	8.59	8.58
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.57	6.52	6.49	6.46	6.43	6.40	6.38	6.36	6.34
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41	5.37	5.33	5.30	5.27	5.24	5.22	5.20	5.18
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.71	4.67	4.63	4.60	4.57	4.54	4.52	4.50	4.48
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.24	4.20	4.16	4.13	4.10	4.08	4.05	4.03	4.02
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.91	3.87	3.83	3.80	3.77	3.74	3.72	3.70	3.68
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.66	3.62	3.58	3.55	3.52	3.50	3.47	3.45	3.44
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.47	3.43	3.39	3.36	3.33	3.30	3.28	3.26	3.24
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.32	3.28	3.24	3.21	3.18	3.15	3.13	3.11	3.09
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.09	3.05	3.01	2.98	2.95	2.92	2.90	2.88	2.86
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	3.01	2.96	2.92	2.89	2.86	2.84	2.81	2.79	2.77

Table 6.2. (Continued)

 $\alpha = 0.05$ v_1

$v_2 \backslash v_1$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.93	2.89	2.85	2.82	2.79	2.76	2.74	2.72	2.70
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.87	2.82	2.79	2.75	2.72	2.70	2.67	2.65	2.63
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.81	2.77	2.73	2.70	2.67	2.64	2.62	2.60	2.58
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.76	2.72	2.68	2.65	2.62	2.59	2.57	2.55	2.53
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72	2.68	2.64	2.60	2.57	2.55	2.52	2.50	2.48
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.68	2.64	2.60	2.56	2.53	2.51	2.48	2.46	2.44
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.65	2.60	2.56	2.53	2.50	2.47	2.45	2.43	2.41
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.62	2.57	2.53	2.50	2.47	2.44	2.42	2.39	2.37
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.59	2.54	2.50	2.47	2.44	2.41	2.39	2.36	2.35
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.56	2.51	2.48	2.44	2.41	2.38	2.36	2.34	2.32
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.54	2.49	2.45	2.42	2.39	2.36	2.34	2.31	2.29
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.51	2.47	2.43	2.39	2.36	2.34	2.31	2.29	2.27
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.49	2.45	2.41	2.37	2.34	2.32	2.29	2.27	2.25
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.48	2.43	2.39	2.36	2.32	2.30	2.27	2.25	2.23
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.46	2.41	2.37	2.34	2.31	2.28	2.26	2.23	2.21
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.13	2.11	2.09
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.22	2.17	2.13	2.09	2.06	2.03	2.01	1.98	1.96
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.10	2.05	2.01	1.98	1.94	1.92	1.89	1.87	1.84
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.99	1.94	1.90	1.87	1.83	1.80	1.78	1.75	1.73

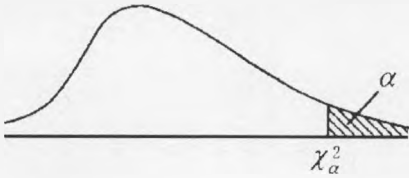
Table 6.2. (Continued)

$\alpha = 0.01$																			
ν_1																			
ν_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	16 211	20 000	21 615	22 500	23 056	23 437	23 715	23 925	24 091	24 224	24 334	24 426	24 505	24 572	24 630	24 681	24 727	24 767	24 803
2	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.5
3	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	43.52	43.39	43.27	43.17	43.08	43.01	42.94	42.88	42.83
4	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	20.82	20.70	20.60	20.51	20.44	20.37	20.31	20.26	20.21
5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.49	13.38	13.29	13.21	13.15	13.09	13.03	12.98	12.94
6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.13	10.03	9.95	9.88	9.81	9.76	9.71	9.66	9.62
7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.27	8.18	8.10	8.03	7.97	7.91	7.87	7.83	7.79
8	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.10	7.01	6.94	6.87	6.81	6.76	6.72	6.68	6.64
9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.31	6.23	6.15	6.09	6.03	5.98	5.94	5.90	5.86
10	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.75	5.66	5.59	5.53	5.47	5.42	5.38	5.34	5.31
11	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.32	5.24	5.16	5.10	5.05	5.00	4.96	4.92	4.89
12	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.99	4.91	4.84	4.77	4.72	4.67	4.63	4.59	4.56
13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.72	4.64	4.57	4.51	4.46	4.41	4.37	4.33	4.30
14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.51	4.43	4.36	4.30	4.25	4.20	4.16	4.12	4.09
15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.52	4.43	4.25	4.18	4.12	4.07	4.02	3.98	3.95	3.91
16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.18	4.10	4.03	3.97	3.92	3.87	3.83	3.80	3.76
17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	4.05	3.97	3.90	3.84	3.79	3.75	3.71	3.67	3.64
18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.94	3.86	3.79	3.73	3.68	3.64	3.60	3.56	3.53
19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.84	3.76	3.70	3.64	3.59	3.54	3.50	3.46	3.43
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.76	3.68	3.61	3.55	3.50	3.46	3.42	3.38	3.35
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.68	3.60	3.54	3.48	3.43	3.38	3.34	3.31	3.27
22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.61	3.54	3.47	3.41	3.39	3.31	3.27	3.24	3.21
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.55	3.47	3.41	3.35	3.30	3.25	3.21	3.18	3.15
24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.50	3.42	3.35	3.30	3.25	3.20	3.16	3.12	3.09

Table 6.2. (Continued)

		$\alpha = 0.01$																		
		ν_1																		
$\nu_2 \backslash$		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
25		9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.45	3.37	3.30	3.25	3.20	3.15	3.11	3.08	3.04
26		9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.40	3.33	3.26	3.20	3.15	3.11	3.07	3.03	3.00
27		9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	3.36	3.28	3.22	3.16	3.11	3.07	3.03	2.99	2.96
28		9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.32	3.25	3.18	3.12	3.07	3.03	2.99	2.95	2.92
29		9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38	3.29	3.21	3.15	3.09	3.04	2.99	2.95	2.92	2.88
30		9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.25	3.18	3.11	3.06	3.01	2.96	2.92	2.89	2.85
40		8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	3.03	2.95	2.89	2.83	2.78	2.74	2.70	2.66	2.63
60		8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.82	2.74	2.68	2.62	2.57	2.53	2.49	2.45	2.42
120		8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.62	2.54	2.48	2.42	2.37	2.33	2.29	2.25	2.22
∞		7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.43	2.39	2.29	2.24	2.19	2.14	2.10	2.06	2.03

Table 7. Critical values of χ^2 distribution (upper tail).



	α								
ν	0.975	0.95	0.50	0.25	0.10	0.05	0.025	0.01	0.001
1	0.001	0.004	0.455	1.32	2.71	3.84	5.02	6.63	10.83
2	0.051	0.103	1.39	2.77	4.61	5.99	7.38	9.21	13.82
3	0.216	0.352	2.37	4.11	6.25	7.82	9.35	11.34	16.27
4	0.484	0.711	3.36	5.39	7.78	9.49	11.14	13.28	18.47
5	0.831	1.15	4.35	6.63	9.24	11.07	12.83	15.09	20.52
6	1.24	1.64	5.35	7.84	10.64	12.59	14.45	16.81	22.46
7	1.69	2.17	6.35	9.04	12.02	14.07	16.01	18.47	24.32
8	2.18	2.73	7.34	10.22	13.36	15.51	17.53	20.09	26.12
9	2.70	3.33	8.34	11.39	14.68	16.92	19.02	21.67	27.88
10	3.25	3.94	9.34	12.55	15.99	18.31	20.48	23.21	29.59
11	3.82	4.57	10.34	13.70	17.27	19.68	21.92	24.72	31.26
12	4.40	5.23	11.34	14.85	18.55	21.03	23.34	26.22	32.91
13	5.01	5.89	12.34	15.98	19.81	22.36	24.74	27.69	34.53
14	5.63	6.57	13.34	17.12	21.06	23.68	26.12	29.14	36.12
15	6.26	7.26	14.34	18.25	22.31	25.00	27.49	30.58	37.70
16	6.91	7.96	15.34	19.37	23.54	26.30	28.85	32.00	39.25
17	7.56	8.67	16.34	20.49	24.77	27.59	30.19	33.41	40.79
18	8.23	9.39	17.34	21.60	25.99	28.87	31.53	34.81	42.31
19	8.91	10.12	18.34	22.72	27.20	30.14	32.85	36.19	43.82
20	9.59	10.85	19.34	23.83	28.41	31.41	34.17	37.57	45.31
21	10.28	11.59	20.34	24.93	29.62	32.67	35.48	38.93	46.80
22	10.98	12.34	21.34	26.04	30.81	33.92	36.78	40.29	48.27
23	11.69	13.09	22.34	27.14	32.01	35.17	38.08	41.64	49.73
24	12.40	13.85	23.34	28.24	33.20	36.42	39.36	42.98	51.18
25	13.12	14.61	24.34	29.34	34.38	37.65	40.65	44.31	52.62
26	13.84	15.38	25.34	30.43	35.56	38.89	41.92	45.64	54.05
27	14.57	16.15	26.34	31.53	36.74	40.11	43.19	46.96	55.48
28	15.31	16.93	27.34	32.62	37.92	41.34	44.46	48.28	56.89
29	16.05	17.71	28.34	33.71	39.09	42.56	45.72	49.59	58.30
30	16.79	18.49	29.34	34.80	40.26	43.77	46.98	50.89	59.70

Table 7. (Continued)

ν	α								
	0.975	0.95	0.50	0.25	0.10	0.05	0.025	0.01	0.001
35	20.57	22.47	34.34	40.22	46.06	49.80	53.20	57.34	66.62
40	24.43	26.51	39.34	45.62	51.81	55.76	59.34	63.69	73.40
45	28.37	30.61	44.34	50.98	57.51	61.66	65.41	69.96	80.08
50	32.36	34.76	49.33	56.33	63.17	67.50	71.42	76.15	86.66
55	36.40	38.96	54.33	61.66	68.80	73.31	77.38	82.29	93.17
60	40.48	43.19	59.33	66.98	74.40	79.08	83.30	88.38	99.61
65	44.60	47.45	64.33	72.28	79.97	84.82	89.18	94.42	105.99
70	48.76	51.74	69.33	77.58	85.53	90.53	95.02	100.43	112.32
75	52.94	56.05	74.33	82.86	91.06	96.22	100.84	106.39	118.60
80	57.15	60.39	79.33	88.13	96.58	101.88	106.63	112.33	124.84
85	61.39	64.75	84.33	93.39	102.08	107.52	112.39	118.24	131.04
90	65.65	69.13	89.33	98.65	107.57	113.15	118.14	124.12	137.21
95	69.92	73.52	94.33	103.90	113.04	118.75	123.86	129.97	143.34
100	74.22	77.93	99.33	109.14	118.50	124.34	129.56	135.81	149.45

When the degree of freedom is more than 100, the critical values of χ^2 can be calculated from $\chi^2 = 0.5[z + \sqrt{2(\nu - 1)}]^2$. Here, z is the upper-tailed value of the standard normal distribution corresponding with given P -value and ν is the degree of freedom.

Table 8. Critical values of the correlation coefficient, r .

ν	Probability α									
	One-tailed Two-tailed	0.25 0.5	0.10 0.20	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.01	0.0025 0.005	0.001 0.002	0.000 0.001
1		0.707	0.951	0.988	0.997	1.000	1.000	1.000	1.000	1.000
2		0.500	0.800	0.900	0.950	0.980	0.990	0.995	0.998	0.999
3		0.404	0.687	0.805	0.878	0.934	0.959	0.974	0.986	0.991
4		0.347	0.608	0.729	0.811	0.882	0.917	0.942	0.963	0.974
5		0.309	0.551	0.669	0.755	0.833	0.875	0.906	0.935	0.951
6		0.281	0.507	0.621	0.707	0.789	0.834	0.870	0.905	0.925
7		0.260	0.472	0.582	0.666	0.750	0.798	0.836	0.875	0.898
8		0.242	0.443	0.549	0.632	0.715	0.765	0.805	0.847	0.872
9		0.228	0.419	0.521	0.602	0.685	0.735	0.776	0.820	0.847
10		0.216	0.398	0.497	0.576	0.658	0.708	0.750	0.795	0.823
11		0.206	0.380	0.476	0.553	0.634	0.684	0.726	0.772	0.801
12		0.197	0.365	0.457	0.532	0.612	0.661	0.703	0.750	0.780
13		0.189	0.351	0.441	0.514	0.592	0.641	0.683	0.730	0.760

Table 8. (Continued)

ν	Probability α									
	One-tailed Two-tailed	0.25 0.5	0.10 0.20	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.01	0.0025 0.005	0.001 0.002	0.000 0.001
14		0.182	0.338	0.426	0.497	0.574	0.623	0.664	0.711	0.742
15		0.176	0.327	0.412	0.482	0.558	0.606	0.647	0.694	0.725
16		0.170	0.317	0.400	0.468	0.542	0.590	0.631	0.678	0.708
17		0.165	0.308	0.389	0.456	0.529	0.575	0.616	0.662	0.693
18		0.160	0.299	0.378	0.444	0.515	0.561	0.602	0.648	0.679
19		0.156	0.291	0.369	0.433	0.503	0.549	0.589	0.635	0.665
20		0.152	0.284	0.360	0.423	0.492	0.537	0.576	0.622	0.652
21		0.148	0.277	0.352	0.413	0.482	0.526	0.565	0.610	0.640
22		0.145	0.271	0.344	0.404	0.472	0.515	0.554	0.599	0.629
23		0.141	0.265	0.337	0.396	0.462	0.505	0.543	0.588	0.618
24		0.138	0.260	0.330	0.338	0.453	0.496	0.534	0.578	0.607
25		0.136	0.255	0.323	0.381	0.445	0.487	0.524	0.568	0.597
26		0.133	0.250	0.317	0.374	0.437	0.479	0.515	0.559	0.588
27		0.131	0.245	0.311	0.367	0.430	0.471	0.507	0.550	0.579
28		0.128	0.241	0.306	0.361	0.423	0.463	0.499	0.541	0.570
29		0.126	0.237	0.301	0.355	0.416	0.456	0.491	0.533	0.562
30		0.124	0.233	0.296	0.349	0.409	0.449	0.484	0.526	0.554
31		0.122	0.229	0.291	0.344	0.403	0.442	0.477	0.518	0.546
32		0.120	0.225	0.287	0.339	0.397	0.436	0.470	0.511	0.539
33		0.118	0.222	0.283	0.334	0.392	0.430	0.464	0.504	0.532
34		0.116	0.219	0.279	0.329	0.386	0.424	0.458	0.498	0.525
35		0.115	0.216	0.275	0.325	0.381	0.418	0.452	0.492	0.519
36		0.113	0.213	0.271	0.320	0.376	0.413	0.446	0.486	0.513
37		0.111	0.210	0.267	0.316	0.371	0.408	0.441	0.480	0.507
38		0.110	0.207	0.264	0.312	0.367	0.403	0.435	0.474	0.501
39		0.108	0.204	0.261	0.308	0.362	0.398	0.430	0.469	0.495
40		0.107	0.202	0.257	0.304	0.358	0.393	0.425	0.463	0.490
41		0.106	0.199	0.254	0.301	0.354	0.389	0.420	0.458	0.484
42		0.104	0.197	0.251	0.297	0.350	0.384	0.416	0.453	0.479
43		0.103	0.195	0.248	0.294	0.346	0.380	0.411	0.449	0.474
44		0.102	0.192	0.246	0.291	0.342	0.376	0.407	0.444	0.469
45		0.101	0.190	0.243	0.288	0.338	0.372	0.403	0.439	0.465
46		0.100	0.188	0.240	0.285	0.335	0.368	0.399	0.435	0.460
47		0.099	0.186	0.238	0.282	0.331	0.365	0.395	0.431	0.456
48		0.098	0.184	0.235	0.279	0.328	0.361	0.391	0.427	0.451
49		0.097	0.182	0.223	0.276	0.325	0.358	0.387	0.423	0.447
50		0.096	0.181	0.231	0.273	0.322	0.354	0.384	0.419	0.443

Table 9. Critical values of the Spearman rank correlation coefficient, r_s .

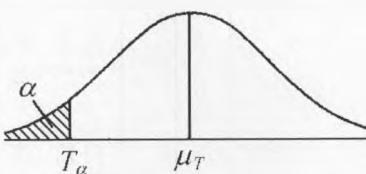
		Probability α								
n	One-tailed	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
	Two-tailed	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
4		0.600	1.000	1.000						
5		0.500	0.800	0.900	1.00	1.000				
6		0.371	0.657	0.829	0.886	0.943	1.000	1.000		
7		0.321	0.571	0.714	0.786	0.893	0.929	0.964	1.000	1.000
8		0.310	0.524	0.643	0.738	0.833	0.881	0.905	0.952	0.976
9		0.267	0.483	0.600	0.700	0.783	0.833	0.867	0.917	0.933
10		0.248	0.455	0.564	0.648	0.745	0.794	0.830	0.879	0.903
11		0.236	0.427	0.536	0.618	0.709	0.755	0.800	0.845	0.873
12		0.217	0.406	0.503	0.587	0.678	0.727	0.769	0.818	0.846
13		0.209	0.385	0.484	0.560	0.648	0.703	0.747	0.791	0.824
14		0.200	0.367	0.464	0.538	0.626	0.679	0.723	0.771	0.802
15		0.189	0.354	0.446	0.521	0.604	0.654	0.700	0.750	0.779
16		0.182	0.341	0.429	0.503	0.582	0.635	0.679	0.729	0.762
17		0.176	0.328	0.414	0.485	0.566	0.615	0.662	0.713	0.748
18		0.170	0.317	0.401	0.472	0.550	0.600	0.643	0.695	0.728
19		0.165	0.309	0.391	0.460	0.535	0.584	0.628	0.677	0.712
20		0.161	0.299	0.380	0.447	0.520	0.570	0.612	0.662	0.696
21		0.156	0.292	0.370	0.435	0.508	0.556	0.599	0.648	0.681
22		0.152	0.284	0.361	0.425	0.496	0.544	0.586	0.634	0.667
23		0.148	0.278	0.353	0.415	0.486	0.532	0.573	0.622	0.654
24		0.144	0.271	0.344	0.406	0.476	0.521	0.562	0.610	0.642
25		0.142	0.265	0.337	0.398	0.466	0.511	0.551	0.598	0.630
26		0.138	0.259	0.331	0.390	0.457	0.501	0.541	0.587	0.619
27		0.136	0.255	0.324	0.382	0.448	0.491	0.531	0.577	0.608

Table 9. (Continued)

n	Probability α									
	One-tailed Two-tailed	0.25 0.50	0.10 0.20	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.01	0.0025 0.005	0.001 0.002	0.0005 0.001
28		0.133	0.250	0.317	0.375	0.440	0.483	0.522	0.567	0.598
29		0.130	0.245	0.312	0.368	0.433	0.475	0.513	0.558	0.589
30		0.128	0.240	0.306	0.362	0.425	0.467	0.504	0.549	0.580
31		0.126	0.236	0.301	0.356	0.418	0.459	0.496	0.541	0.571
32		0.124	0.232	0.296	0.350	0.412	0.452	0.489	0.533	0.563
33		0.121	0.229	0.291	0.345	0.405	0.446	0.482	0.525	0.554
34		0.120	0.225	0.287	0.340	0.399	0.439	0.475	0.517	0.547
35		0.118	0.222	0.283	0.335	0.394	0.433	0.468	0.510	0.539
36		0.116	0.219	0.279	0.330	0.388	0.427	0.462	0.504	0.533
37		0.114	0.216	0.275	0.325	0.382	0.421	0.456	0.497	0.526
38		0.113	0.212	0.271	0.321	0.378	0.415	0.450	0.491	0.519
39		0.111	0.210	0.267	0.317	0.373	0.410	0.444	0.485	0.513
40		0.110	0.207	0.264	0.313	0.368	0.405	0.439	0.479	0.507
41		0.108	0.204	0.261	0.309	0.364	0.400	0.433	0.473	0.501
42		0.107	0.202	0.257	0.305	0.359	0.395	0.428	0.468	0.495
43		0.105	0.199	0.254	0.301	0.355	0.391	0.423	0.463	0.490
44		0.104	0.197	0.251	0.298	0.351	0.386	0.419	0.458	0.484
45		0.103	0.194	0.248	0.294	0.347	0.382	0.414	0.453	0.479
46		0.102	0.192	0.246	0.291	0.343	0.378	0.410	0.448	0.474
47		0.101	0.190	0.243	0.288	0.340	0.374	0.405	0.443	0.469
48		0.100	0.188	0.240	0.285	0.336	0.370	0.401	0.439	0.465
49		0.098	0.186	0.238	0.282	0.333	0.366	0.397	0.434	0.460
50		0.097	0.184	0.235	0.279	0.329	0.363	0.393	0.430	0.456

Table 10. Critical values of the $T_{\alpha(n)}$ in Wilcoxon rank sum test.

(n is the valid pairs)



One-tailed α					One-tailed α				
0.05 0.025 0.01 0.005					0.05 0.025 0.01 0.005				
Two-tailed α					Two-tailed α				
n	0.10	0.05	0.02	0.01	n	0.10	0.05	0.02	0.01
5	1				28	130	116	101	91
6	1	1			29	140	126	110	100
7	3	2	0		30	151	137	120	109
8	5	3	1	0	31	163	147	130	118
9	8	5	3	1	32	175	159	140	128
10	10	8	5	3	33	187	170	151	138
11	13	10	7	5	34	200	182	162	148
12	17	13	9	7	35	213	195	173	159
13	21	17	12	9	36	227	208	185	171
14	25	21	15	12	37	241	221	198	182
15	30	25	19	15	38	256	235	211	194
16	35	29	23	19	39	271	249	224	207
17	41	34	27	23	40	286	264	238	220
18	47	40	32	27	41	302	279	252	233
19	53	46	37	32	42	319	294	266	247
20	60	52	43	37	43	336	301	281	261
21	67	58	49	42	44	353	327	296	276
22	75	65	55	48	45	371	343	312	291
23	83	73	62	54	46	389	361	328	307
24	91	81	69	61	47	407	378	345	322
25	100	89	76	68	48	426	396	362	339
26	110	98	84	75	49	446	415	379	355
27	119	107	93	83	50	466	434	397	373

Table 11. Critical values of T_{α}^L and T_{α}^U for Wincoxon–Mann–Whitney rank sum test.

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Table 11. (Continued)

n_1 (smaller n)	$n_2 - n_1$										
	0	1	2	3	4	5	6	7	8	9	10
6	28-50	29-55	31-59	33-63	35-67	37-71	38-76	40-80	42-84	44-88	46-92
	26-52	27-57	29-61	31-65	32-70	34-74	35-79	37-83	38-88	40-92	42-96
	24-54	25-59	27-63	28-68	29-73	30-78	32-82	33-87	34-92	36-96	37-101
	23-55	24-60	25-65	26-70	27-75	28-80	30-84	31-89	32-94	33-99	34-104
7	39-66	41-71	43-76	45-81	47-86	49-91	52-95	54-100	56-105	58-110	61-114
	36-69	38-74	40-79	42-84	44-89	46-94	48-99	50-104	52-109	54-114	56-119
	34-71	35-77	37-82	39-87	40-93	42-98	44-103	45-109	47-114	49-119	51-124
	32-73	34-78	35-84	37-89	38-95	40-100	41-106	43-111	44-117	45-122	47-128
8	51-85	54-90	56-96	59-101	62-106	64-112	67-117	69-123	72-128	75-133	77-139
	49-87	51-93	53-99	55-105	58-110	60-116	63-121	65-127	67-133	70-138	72-144
	45-91	47-97	49-103	51-109	53-115	56-120	58-126	60-132	62-138	64-144	66-150
	43-93	45-99	47-105	49-111	51-117	53-123	54-130	56-136	58-142	60-148	62-154
9	66-105	69-111	72-117	75-123	78-129	81-135	84-141	87-147	90-153	93-159	96-165
	63-108	65-115	68-121	71-127	73-134	76-140	79-146	82-152	84-159	87-165	90-171
	59-112	61-119	63-126	66-132	68-139	71-145	73-152	76-158	78-165	81-171	83-178
	56-115	58-122	61-128	63-135	65-142	67-149	70-155	72-162	74-169	76-176	78-183
10	82-128	86-134	89-141	92-148	96-154	99-161	103-167	106-174	110-180	113-187	117-193
	78-132	81-139	85-145	88-152	91-159	94-166	97-173	100-180	103-187	107-193	110-200
	74-136	77-143	79-151	82-158	85-165	88-172	91-179	93-187	96-194	99-201	102-208
	71-139	74-146	76-154	79-161	81-169	84-176	86-184	89-191	92-198	94-206	97-213

Table 12. Critical values of q for multiple comparison.

$\alpha = 0.05$																				
$\nu \backslash k$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	$k \backslash \nu$
1	17.97	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07	50.59	51.96	53.20	54.33	55.36	56.32	57.22	53.04	58.83	59.56	1
2	6.08	8.33	9.80	10.88	11.74	12.44	13.03	13.54	13.99	14.39	14.75	15.08	15.38	15.65	15.91	16.14	16.37	16.57	16.77	2
3	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	9.72	9.95	10.15	10.35	10.52	10.69	10.84	10.98	11.11	11.24	3
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	8.03	8.21	8.37	8.52	8.66	8.79	8.91	9.03	9.13	9.23	4
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21	5
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65	6.79	6.92	7.03	7.14	7.24	7.34	7.43	7.51	7.59	6
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30	6.43	6.55	6.66	6.76	6.85	6.94	7.02	7.10	7.17	7
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87	8
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87	5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.64	9
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72	5.83	5.93	6.03	6.11	6.19	6.27	6.34	6.40	6.47	10
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.90	5.98	6.06	6.13	6.20	6.27	6.33	11
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51	5.61	5.71	5.80	5.88	5.95	6.02	6.09	6.15	6.21	12
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43	5.53	5.63	5.71	5.79	5.86	5.93	5.99	6.05	6.11	13
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.55	5.64	5.71	5.79	5.85	5.91	5.97	6.03	14
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31	5.40	5.49	5.57	5.65	5.72	5.78	5.85	5.90	5.96	15
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26	5.35	5.44	5.52	5.59	5.66	5.73	5.79	5.84	5.90	16
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21	5.31	5.39	5.47	5.54	5.61	5.67	5.73	5.79	5.84	17
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17	5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79	18
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14	5.23	5.31	5.39	5.46	5.53	5.59	5.65	5.70	5.75	19
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.36	5.43	5.49	5.55	5.61	5.66	5.71	20
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10	5.18	5.25	5.32	5.38	5.44	5.49	5.55	5.59	24
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92	5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.47	30
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82	4.90	4.98	5.04	5.11	5.15	5.22	5.27	5.31	5.36	40
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81	4.88	4.94	5.00	5.06	5.11	5.15	5.20	5.24	60
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64	4.71	4.78	4.84	4.90	4.95	5.00	5.04	5.09	5.13	120
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	4.62	4.68	4.74	4.80	4.85	4.89	4.93	4.97	5.01	∞

Table 12. (Continued)

$\alpha = 0.01$																					
$\nu \backslash k$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	$k \backslash \nu$	
1	90.03	135.0	164.3	185.6	202.2	215.8	227.2	237.0	245.6	253.2	260.0	266.2	271.8	277.0	281.8	286.3	290.4	294.3	298.0	1	
2	14.04	19.02	22.29	24.72	26.63	28.20	29.53	30.68	31.69	32.59	33.40	34.13	34.81	35.43	36.00	36.53	37.03	37.50	37.95	2	
3	8.26	10.62	12.17	13.33	14.24	15.00	15.64	16.20	16.69	17.13	17.53	17.89	18.22	18.52	18.81	19.07	19.32	19.55	19.77	3	
4	6.51	8.12	9.17	9.96	10.58	11.10	11.55	11.93	12.27	12.57	12.84	13.09	13.32	13.53	13.73	13.91	14.08	14.24	14.40	4	
5	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48	10.70	10.89	11.08	11.24	11.40	11.55	11.68	11.81	11.93	5	
6	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30	9.48	9.65	9.81	9.95	10.08	10.21	10.32	10.43	10.54	6	
7	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55	8.71	8.86	9.00	9.12	9.24	9.35	9.46	9.55	9.65	7	
8	4.75	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86	8.03	8.18	8.31	8.44	8.55	8.66	8.76	8.85	8.94	9.03	8	
9	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49	7.65	7.78	7.91	8.03	8.13	8.23	8.33	8.41	8.49	8.57	9	
10	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36	7.49	7.60	7.71	7.81	7.91	7.99	8.08	8.15	8.23	10	
11	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13	7.25	7.36	7.46	7.56	7.65	7.73	7.81	7.88	7.95	11	
12	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94	7.06	7.17	7.26	7.36	7.44	7.52	7.59	7.66	7.73	12	
13	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79	6.90	7.01	7.10	7.19	7.27	7.35	7.42	7.48	7.55	13	
14	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66	6.77	6.87	6.96	7.05	7.13	7.20	7.27	7.33	7.39	14	
15	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55	6.66	6.70	6.84	6.98	7.00	7.07	7.14	7.20	7.26	15	
16	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46	6.56	6.66	6.74	6.82	6.90	6.97	7.03	7.09	7.15	16	
17	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38	6.48	6.57	6.66	6.73	6.81	6.87	6.94	7.00	7.05	17	
18	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31	6.41	6.50	6.58	6.65	6.73	6.79	6.85	6.91	6.97	18	
19	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25	6.34	6.43	6.51	6.58	6.65	6.72	6.78	6.84	6.89	19	
20	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19	6.28	6.37	6.45	6.52	6.59	6.65	6.71	6.77	6.82	20	
24	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02	6.11	6.19	6.26	6.33	6.39	6.45	6.51	6.56	6.61	24	
30	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85	5.93	6.01	6.08	6.14	6.20	6.26	6.31	6.36	6.41	30	
40	3.82	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60	5.69	5.76	5.83	5.90	5.96	6.02	6.07	6.12	6.16	6.21	40	
60	3.76	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45	5.53	5.60	5.67	5.73	5.78	5.84	5.89	5.93	5.97	6.01	60	
120	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.37	5.44	5.50	5.56	5.61	5.66	5.71	5.75	5.79	5.83	120	
∞	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23	5.29	5.35	5.40	5.45	5.49	5.54	5.57	5.61	5.65	∞	

Table 13.1. Critical values for Dunnett-*t* test (one tailed).
(The upper line is for $\alpha = 0.05$, the lower line is for $\alpha = 0.01$.)

DF of the error (ν)	Number of treatments (despite the control) <i>T</i>								
	1	2	3	4	5	6	7	8	9
5	2.02	2.44	2.68	2.85	2.98	3.08	3.16	3.24	3.30
	3.37	3.90	4.21	4.43	4.60	4.73	4.85	4.94	5.03
6	1.94	2.34	2.56	2.71	2.83	2.92	3.00	3.07	3.12
	3.14	3.61	3.88	4.07	4.21	4.33	4.43	4.51	4.59
7	1.89	2.27	2.48	2.62	2.73	2.82	2.89	2.95	3.01
	3.00	3.42	3.66	3.83	3.96	4.07	4.15	4.23	4.30
8	1.86	2.22	2.42	2.55	2.66	2.74	2.81	2.87	2.92
	2.90	3.29	3.51	3.67	3.79	3.88	3.96	4.03	4.09
9	1.83	2.18	2.37	2.50	2.60	2.68	2.75	2.81	2.86
	2.82	3.19	3.40	3.55	3.66	3.75	3.82	3.89	3.94
10	1.81	2.15	2.34	2.47	2.56	2.64	2.70	2.76	2.81
	2.76	3.11	3.31	3.45	3.56	3.64	3.71	3.78	3.83
11	1.80	2.13	2.31	2.44	2.53	2.60	2.67	2.72	2.77
	2.72	3.06	3.25	3.38	3.48	3.56	3.63	3.69	3.74
12	1.78	2.11	2.29	2.41	2.50	2.58	2.64	2.69	2.74
	2.68	3.01	3.19	3.32	3.42	3.50	3.56	3.62	3.67
13	1.77	2.09	2.27	2.39	2.48	2.55	2.61	2.66	2.71
	2.65	2.97	3.15	3.27	3.37	3.44	3.51	3.56	3.61
14	1.76	2.08	2.25	2.37	2.46	2.53	2.59	2.64	2.69
	2.62	2.94	3.11	3.23	3.32	3.40	3.46	3.51	3.56
15	1.75	2.07	2.24	2.36	2.44	2.51	2.57	2.62	2.67
	2.60	2.91	3.08	3.20	3.29	3.36	3.42	3.47	3.52
16	1.75	2.06	2.23	2.34	2.43	2.50	2.56	2.61	2.65
	2.58	2.88	3.05	3.17	3.26	3.33	3.39	3.44	3.48
17	1.74	2.05	2.22	2.33	2.42	2.49	2.54	2.59	2.64
	2.57	2.86	3.03	3.14	3.23	3.30	3.36	3.41	3.45
18	1.73	2.04	2.21	2.32	2.41	2.48	2.53	2.58	2.62
	2.55	2.84	3.01	3.12	3.21	3.27	3.33	3.38	3.42
19	1.73	2.03	2.20	2.31	2.40	2.47	2.52	2.57	2.61
	2.54	2.83	2.99	3.10	3.18	3.25	3.31	3.36	3.40
20	1.72	2.03	2.19	2.30	2.39	2.46	2.51	2.56	2.60
	2.53	2.81	2.97	3.08	3.17	3.23	3.29	3.34	3.38
24	1.71	2.01	2.17	2.28	2.36	2.43	2.48	2.53	2.57
	2.49	2.77	2.92	3.03	3.11	3.17	3.22	3.27	3.31

Table 13.1. (Continued)

DF of the error (ν)	Number of treatments (despite the control) T								
	1	2	3	4	5	6	7	8	9
30	1.70	1.99	2.15	2.25	2.33	2.40	2.45	2.50	2.54
	2.46	2.72	2.87	2.97	3.05	3.11	3.16	3.21	3.24
40	1.68	1.97	2.13	2.23	2.31	2.37	2.42	2.47	2.51
	2.42	2.68	2.82	2.92	2.99	3.05	3.10	3.14	3.18
60	1.67	1.95	2.10	2.21	2.28	2.35	2.39	2.44	2.48
	2.39	2.64	2.78	2.87	2.94	3.00	3.04	3.08	3.12
120	1.66	1.93	2.08	2.18	2.26	2.32	2.37	2.41	2.45
	2.36	2.60	2.73	2.82	2.89	2.94	2.99	3.03	3.06
∞	1.64	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42
	2.33	2.56	2.68	2.77	2.84	2.89	2.93	2.97	3.00

Table 13.2. Critical values of the Dunnett- t test (two tailed).
(The upper line is for $\alpha = 0.05$, the lower line is for $\alpha = 0.01$.)

DF of the error (ν)	Number of treatments (despite the control) T								
	1	2	3	4	5	6	7	8	9
5	2.57	3.03	3.29	3.48	3.62	3.73	3.82	3.90	3.97
	4.03	4.63	4.98	5.22	5.41	5.56	5.69	5.80	5.89
6	2.45	2.86	3.10	3.26	3.39	3.49	3.57	3.64	3.71
	3.71	4.21	4.51	4.70	4.87	5.00	5.10	5.20	5.28
7	2.36	2.75	2.97	3.12	3.24	3.33	3.41	3.47	3.53
	3.50	3.95	4.21	4.39	4.53	4.64	4.74	4.82	4.89
8	2.31	2.67	2.88	3.02	3.13	3.22	3.29	3.35	3.41
	3.36	3.77	4.00	4.17	4.29	4.40	4.48	4.56	4.62
9	2.26	2.61	2.81	2.95	3.05	3.14	3.20	3.26	3.32
	3.25	3.63	3.85	4.01	4.12	4.22	4.30	4.37	4.43
10	2.23	2.57	2.76	2.89	2.99	3.07	3.14	3.19	3.24
	3.17	3.53	3.74	3.88	3.99	4.08	4.16	4.22	4.28
11	2.20	2.53	2.72	2.84	2.94	3.02	3.08	3.14	3.19
	3.11	3.45	3.65	3.79	3.88	3.98	4.05	4.11	4.16

Table 13.2. (Continued)

DF of the error (ν)	Number of treatments (despite the control) T								
	1	2	3	4	5	6	7	8	9
12	2.18	2.50	2.68	2.81	2.90	2.98	3.04	3.05	3.14
	3.05	3.39	3.58	3.71	3.81	3.89	3.96	4.02	4.07
13	2.16	2.48	2.65	2.78	2.87	2.94	3.00	3.06	3.10
	3.01	3.33	3.52	3.65	3.74	3.82	3.89	3.94	3.99
14	2.14	2.46	2.63	2.75	2.84	2.91	2.97	3.02	3.07
	2.98	3.29	3.47	3.59	3.69	3.76	3.83	3.88	3.93
15	2.13	2.44	2.61	2.73	2.82	2.89	2.95	3.00	3.04
	2.95	3.25	3.43	3.55	3.64	3.71	3.78	3.83	3.88
16	2.12	2.42	2.59	2.71	2.80	2.87	2.92	2.97	3.02
	2.92	3.22	3.39	3.51	3.60	3.67	3.73	3.78	3.83
17	2.11	2.41	2.58	2.69	2.78	2.85	2.90	2.95	3.00
	2.90	3.19	3.36	3.47	3.56	3.63	3.69	3.74	3.79
18	2.10	2.40	2.56	2.68	2.76	2.83	2.89	2.94	2.98
	2.88	3.17	3.33	3.44	3.53	3.60	3.66	3.71	3.75
19	2.09	2.39	2.55	2.66	2.75	2.81	2.87	2.92	2.96
	2.86	3.15	3.31	3.42	3.50	3.57	3.63	3.68	3.72
20	2.09	2.38	2.54	2.65	2.73	2.80	2.86	2.90	2.95
	2.85	3.13	3.29	3.40	3.48	3.55	3.60	3.65	3.69
24	2.06	2.35	2.51	2.61	2.70	2.76	2.81	2.86	2.90
	2.80	3.07	3.22	3.32	3.40	3.47	3.52	3.57	3.61
30	2.04	2.32	2.47	2.58	2.66	2.72	2.77	2.82	2.86
	2.75	3.01	3.15	3.25	3.33	3.39	3.44	3.49	3.52
40	2.02	2.29	2.44	2.54	2.62	2.68	2.73	2.77	2.81
	2.70	2.95	3.09	3.19	3.26	3.32	3.37	3.41	3.44
60	2.00	2.27	2.41	2.51	2.58	2.64	2.69	2.73	2.77
	2.66	2.90	3.03	3.12	3.19	3.25	3.29	3.33	3.37
120	1.98	2.24	2.38	2.47	2.55	2.60	2.65	2.69	2.73
	2.62	2.84	2.97	3.06	3.12	3.18	3.22	3.26	3.29
∞	1.96	2.21	2.35	2.44	2.51	2.57	2.61	2.65	2.69
	2.58	2.79	2.92	3.00	3.06	3.11	3.15	3.19	3.22

Table 14. Sample size for sampling survey on mean.

$\alpha = 0.05$										
σ/δ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	7	8	9	9	11	12	13	14	15	17
2	18	20	22	23	25	27	29	31	33	35
3	38	40	42	45	47	50	53	56	58	61
4	64	68	71	74	77	81	84	88	91	95
5	99	103	107	111	115	119	123	128	132	137
6	141	146	151	156	160	165	170	176	181	186
7	191	196	202	207	213	219	225	231	237	243
8	249	255	261	268	274	281	288	294	301	308
9	315	322	329	336	343	351	358	366	373	381
10	389	396	404	412	420	428	437	445	453	462
11	470	478	487	496	505	514	523	532	541	550
12	559	569	578	588	597	607	617	626	636	646
13	656	667	677	687	697	708	718	729	740	750
14	761	772	783	794	805	816	828	839	851	862
15	874	885	897	909	921	933	945	957	969	982
16	994	1 006	1 019	1 032	1 044	1 057	1 070	1 083	1 096	1 109
17	1 122	1 135	1 149	1 162	1 175	1 189	1 203	1 216	1 230	1 244
18	1 258	1 272	1 286	1 300	1 314	1 329	1 343	1 358	1 372	1 387
19	1 402	1 416	1 431	1 446	1 461	1 476	1 491	1 507	1 522	1 537
20	1 553	1 568	1 583	1 600	1 616	1 631	1 647	1 663	1 680	1 696

$\alpha = 0.01$										
1	11	12	14	15	17	19	21	23	26	28
2	31	34	36	39	43	46	49	53	56	60
3	64	68	72	77	81	86	90	95	100	105
4	110	116	121	127	133	139	145	151	157	164
5	170	177	184	191	198	205	213	220	228	235
6	243	251	260	268	277	285	294	303	312	321
7	331	340	350	360	370	380	390	400	411	421
8	432	443	454	465	476	487	499	511	522	534
9	546	559	571	583	596	609	622	635	648	661
10	674	688	702	715	729	743	758	772	787	801
11	816	831	846	861	876	892	907	923	939	955
12	971	987	1 004	1 020	1 037	1 054	1 070	1 087	1 105	1 122
13	1 139	1 157	1 175	1 193	1 211	1 229	1 247	1 265	1 284	1 303
14	1 321	1 340	1 359	1 379	1 398	1 417	1 437	1 457	1 477	1 497
15	1 517	1 537	1 558	1 578	1 599	1 620	1 641	1 662	1 683	1 704
16	1 726	1 747	1 769	1 791	1 813	1 835	1 858	1 880	1 903	1 925
17	1 948	1 971	1 994	2 017	2 041	2 064	2 088	2 112	2 136	2 160
18	2 184	2 208	2 232	2 257	2 282	2 307	2 332	2 357	2 382	2 408
19	2 433	2 459	2 485	2 511	2 537	2 563	2 589	2 616	2 643	2 669
20	2 696	2 723	2 750	2 778	2 805	2 833	2 860	2 888	2 916	2 943

Table 15. Sample size for sampling survey on probability.

$\alpha = 0.05$										
p										
$\delta \backslash$	0.50	0.45 0.55	0.40 0.60	0.35 0.65	0.30 0.70	0.25 0.75	0.20 0.80	0.15 0.85	0.10 0.90	0.05 0.95
0.200	24	24	23	22	20	18	15			
0.180	30	29	28	27	25	22	19			
0.160	38	37	36	34	32	28	24			
0.140	49	49	47	45	41	37	31	25		
0.120	67	66	64	61	56	50	43	34		
0.100	96	95	92	87	81	72	61	49		
0.090	119	117	114	108	100	89	76	60	43	
0.080	150	149	144	137	126	113	96	77	54	
0.070	196	194	188	178	165	147	125	100	71	
0.060	267	264	256	243	224	200	171	136	96	
0.050	384	380	369	350	323	288	246	196	138	73
0.045	474	470	455	432	498	356	304	242	171	90
0.040	600	594	576	546	504	450	384	306	216	114
0.035	784	776	753	713	659	588	502	400	282	149
0.030	1 067	1 056	1 024	971	896	800	683	544	384	203
0.025	1 537	1 521	1 475	1 398	1 291	1 152	983	784	553	292
0.020	2 401	2 377	2 305	2 185	2 017	1 801	1 537	1 225	864	456
0.015	4 268	4 226	4 098	3 884	3 585	3 201	2 732	2 177	1 537	811
0.010	9 604	9 508	9 220	8 740	8 067	7 203	6 147	4 898	3 457	1 825
0.005	38 415	38 031	36 878	34 958	32 269	28 811	24 586	19 592	13 830	7 299

$\alpha = 0.01$										
0.200	41	41	40	38	35	31	27			
0.180	51	51	49	47	43	38	33			
0.160	65	64	62	59	54	49	41			
0.140	85	84	81	77	71	63	54	43		
0.120	115	114	111	105	97	86	74	59		
0.100	166	164	159	151	139	124	106	85		
0.090	205	203	197	186	172	154	131	104	74	
0.080	259	257	249	236	218	194	166	132	93	
0.070	339	335	325	308	284	254	217	173	122	
0.060	461	456	442	419	387	346	295	235	166	
0.050	664	657	637	604	557	498	425	338	239	125
0.045	819	811	786	746	688	614	524	418	295	156
0.040	1 037	1 026	995	944	871	778	664	529	373	197
0.035	1 354	1 341	1 300	1 232	1 138	1 016	867	691	488	257
0.030	1 843	1 825	1 770	1 677	1 548	1 382	1 180	940	664	350
0.025	2 654	2 628	2 548	2 415	2 230	1 991	1 699	1 354	956	504
0.020	4 147	4 106	3 981	3 774	3 484	3 111	2 654	2 115	1 493	788
0.015	7 373	7 299	7 078	6 710	6 193	5 530	4 719	3 760	2 654	1 401
0.010	16 588	16 422	15 924	15 095	13 934	12 441	10 616	8 460	5 972	3 152
0.005	66 349	65 686	63 696	60 378	55 734	49 762	42 464	33 838	23 886	12 607

Table 16. Sample size for comparison between sample mean and a given constant by t -test.

Two-sided test One-sided test	$\alpha = 0.01$					$\alpha = 0.02$					$\alpha = 0.05$					$\alpha = 0.1$				
	$\alpha = 0.005$					$\alpha = 0.01$					$\alpha = 0.025$					$\alpha = 0.05$				
	β					β					β					β				
δ/σ	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5
0.05																				
0.10																				
0.15																				122
0.20										139					98					69
0.25					110					90				128	64			139	101	45
0.30				134	78				115	63			119	90	45		122	97	71	32
0.35			125	99	58			109	85	47		109	88	67	34		90	72	52	24
0.40		115	97	77	45		102	85	66	37	117	84	68	52	26	100	70	55	41	19
0.45		92	77	62	37	110	81	68	53	30	93	67	54	41	21	80	55	44	32	15
0.50	100	75	63	51	30	90	66	55	43	25	76	54	44	34	18	65	45	36	27	13
0.55	83	63	53	42	26	75	55	46	36	21	63	45	37	28	15	54	38	30	22	11
0.60	71	53	45	36	22	63	47	39	31	18	54	39	32	24	13	46	32	26	19	9
0.65	61	46	39	32	20	55	41	34	27	16	46	33	27	21	12	39	28	22	17	8
0.70	53	40	34	28	17	47	35	30	24	14	40	29	24	19	10	34	24	19	15	8
0.75	47	36	30	25	16	42	31	26	21	13	35	26	21	16	9	30	21	17	13	7
0.80	41	32	27	22	14	37	28	24	19	12	31	23	19	15	9	27	19	15	12	6
0.85	37	29	24	20	13	33	25	21	17	11	28	21	17	13	8	24	17	14	11	6
0.90	34	26	22	18	12	30	23	19	16	10	25	19	16	12	7	21	15	13	10	5
0.95	31	24	20	17	11	27	21	18	14	9	23	17	14	11	7	19	14	11	9	5
1.00	28	22	19	16	10	25	19	16	13	9	21	16	13	10	6	18	13	11	8	5

Table 17. Sample size for comparison between two means by *t*-test.

Two-sided test One-sided test	$\alpha = 0.01$					$\alpha = 0.02$					$\alpha = 0.05$					$\alpha = 0.1$				
	$\alpha = 0.005$					$\alpha = 0.01$					$\alpha = 0.025$					$\alpha = 0.05$				
δ/σ	β					β					β					β				
	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5
0.05																				
0.10																				
0.15																				
0.20																				136
0.25															124					88
0.30										123					87					61
0.35					110					90					64				102	45
0.40					85					70					100				108	78
0.45				118	68					101					79				86	62
0.50				96	55			106		82					64			108	70	51
0.55			101	79	46		106	88	68	38			105	86	32			88	58	42
0.60			101	85	67	39		89	74	58	32		87	71	53	27	105	73	54	36
0.65		101	87	73	57	34	104	77	63	49	28	104	88	63	45	23	89	61	49	30
0.70	100	75	63	50	29	90	66	55	43	24	76	55	44	34	17	66	45	36	26	12
0.75	88	66	55	44	26	79	58	48	38	21	67	48	39	29	15	57	40	32	23	11
0.80	77	58	49	39	23	70	51	43	33	19	59	42	34	26	14	50	35	28	21	10
0.85	69	52	43	35	21	62	46	38	30	17	52	37	31	23	12	45	31	25	18	9
0.90	62	46	39	31	19	55	41	34	27	15	47	34	27	21	11	40	28	22	16	8
0.95	55	42	35	28	17	50	37	31	24	14	42	30	25	19	10	36	25	20	15	7
1.00	50	38	32	26	15	45	33	28	22	13	38	27	23	17	9	33	23	18	14	7

Table 17. (Continued)

Two-sided test One-sided test	$\alpha = 0.01$					$\alpha = 0.02$					$\alpha = 0.05$					$\alpha = 0.1$				
	$\alpha = 0.005$					$\alpha = 0.01$					$\alpha = 0.025$					$\alpha = 0.05$				
	β					β					β					β				
δ/σ	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5
1.1	42	32	27	22	13	38	28	23	19	11	32	23	19	14	8	27	19	15	11	6
1.2	36	27	23	18	11	32	24	20	16	9	27	20	16	12	7	23	16	13	10	5
1.3	31	23	20	16	10	28	21	17	14	8	23	17	14	11	6	20	14	11	9	5
1.4	27	20	17	14	9	24	18	15	12	8	20	15	12	10	6	17	12	10	8	4
1.5	24	18	15	13	8	21	16	14	11	7	18	13	11	9	5	15	11	9	7	4
1.6	21	16	14	11	7	19	14	12	10	6	16	12	10	8	5	14	10	8	6	4
1.7	19	15	13	10	7	17	13	11	9	6	14	11	9	7	4	12	9	7	6	3
1.8	17	13	11	10	6	15	12	10	8	5	13	10	8	6	4	11	8	7	5	
1.9	16	12	11	9	6	14	11	9	8	5	12	9	7	6	4	10	7	6	5	
2.0	14	11	10	8	6	13	10	9	7	5	11	8	7	6	4	9	7	6	4	
2.1	13	10	9	8	5	12	9	8	7	5	10	8	6	5	3	8	6	5	4	
2.2	12	10	8	7	5	11	9	7	6	4	9	7	6	5		8	6	5	4	
2.3	11	9	8	7	5	10	8	7	6	4	9	7	6	5		7	5	5	4	
2.4	11	9	8	6	5	10	8	7	6	4	8	6	5	4		7	5	4	4	
2.5	10	8	7	6	4	9	7	6	5	4	8	6	5	4		6	5	4	3	
3.0	8	6	6	5	4	7	6	5	4	3	6	5	4	4		5	4	3		
3.5	6	5	5	4	3	6	5	4	4		5	4	4	3		4	3			
4.0	6	5	4	4		5	4	4	3		4	4	3			4				

Table 18.1. Sample size for comparison between two sample frequencies (one tailed).

Line 1: $\alpha = 0.05$, $1 - \beta = 0.80$

Line 2: $\alpha = 0.05$, $1 - \beta = 0.90$

Line 3: $\alpha = 0.01$, $1 - \beta = 0.95$

The smaller rate (%)	Difference of the two rates (%), δ													
	5	10	15	20	25	30	35	40	45	50	55	60	65	70
5	330	105	55	35	25	20	16	13	11	9	8	7	6	6
	460	145	76	48	34	26	21	17	15	13	11	9	8	7
	850	270	140	89	63	47	37	30	25	21	19	17	14	13
10	540	155	76	47	32	23	19	15	13	11	9	8	7	6
	740	210	105	64	44	33	25	21	17	14	12	11	9	8
	1370	390	195	120	81	60	46	37	30	25	21	19	16	14
15	710	200	94	56	38	27	21	17	14	12	10	8	7	6
	990	270	130	77	52	38	29	22	19	16	13	10	10	8
	1820	500	240	145	96	69	52	41	33	27	22	20	17	14
20	860	230	110	63	42	30	22	18	15	12	10	8	7	6
	1190	320	150	88	58	41	31	24	20	16	14	11	10	8
	2190	590	280	160	105	76	57	44	35	28	23	20	17	14
25	980	260	120	69	45	32	24	19	15	12	10	8	7	
	1360	360	165	96	63	44	33	25	21	16	14	11	9	
	2510	660	300	175	115	81	60	46	36	29	23	20	16	
30	1080	280	130	73	47	33	24	19	15	12	10	8		
	1500	390	175	100	65	46	33	25	21	16	13	11		
	2760	720	230	185	120	84	61	47	36	28	22	19		
35	1160	300	135	75	48	33	24	19	15	12	9			
	1600	410	185	105	67	46	33	25	20	16	12			
	2960	750	340	190	125	85	61	46	35	27	21			
40	1210	310	135	76	48	33	24	18	14	11				
	1670	420	190	105	67	46	33	24	19	14				
	3080	780	350	195	125	84	60	44	33	25				
45	1230	310	135	75	47	32	22	17	13					
	1710	430	190	105	65	44	31	22	17					
	3140	790	350	190	120	81	57	41	30					
50	1230	310	135	73	45	30	21	15						
	1710	420	185	100	63	41	29	21						
	3140	780	340	185	115	76	52	37						

Table 18.2. Sample size for comparison between two sample frequencies (two tailed).

Line 1: $\alpha = 0.05, 1 - \beta = 0.80$

Line 2: $\alpha = 0.05, 1 - \beta = 0.90$

Line 3: $\alpha = 0.01, 1 - \beta = 0.95$

The smaller rate (%)	Difference of the two rates (%), δ													
	5	10	15	20	25	30	35	40	45	50	55	60	65	70
5	420	130	69	44	31	24	20	16	14	12	10	9	9	7
	570	175	93	59	42	32	25	21	18	15	13	11	10	9
	960	300	155	100	71	54	42	34	28	24	21	19	16	14
10	680	195	96	59	41	30	23	19	16	13	11	10	9	7
	910	260	130	79	54	40	31	24	21	18	15	13	11	10
	1550	440	220	135	92	68	52	41	34	28	23	21	18	15
15	910	250	120	71	48	34	26	21	17	14	12	10	9	8
	1220	330	160	95	64	46	35	27	22	19	16	13	11	10
	2060	560	270	160	110	78	59	47	37	31	25	21	19	16
20	1090	290	135	80	53	38	28	22	18	15	13	10	9	7
	1460	390	185	105	71	51	38	29	23	20	16	14	11	10
	2470	660	310	180	120	86	64	50	40	32	26	21	19	15
25	1250	330	150	88	57	40	30	23	19	15	13	10	9	
	1680	440	200	115	77	54	40	31	24	20	16	13	11	
	2840	740	340	200	130	92	68	52	41	32	26	21	18	
30	1380	360	160	93	60	42	31	23	19	15	12	10		
	1840	480	220	125	80	56	41	31	24	20	16	13		
	3120	810	370	210	135	95	69	53	41	32	25	21		
35	1470	380	170	96	61	42	31	23	18	14	11			
	1970	500	225	130	82	57	41	31	23	19	15			
	3340	850	380	215	140	96	69	52	40	31	23			
40	1530	390	175	97	61	42	30	22	17	13				
	2050	520	230	130	82	56	40	29	22	18				
	3480	880	390	220	140	95	68	50	37	28				
45	1560	390	175	96	60	40	28	21	16					
	2100	520	230	130	80	54	38	27	21					
	3550	890	390	215	135	92	64	47	34					
50	1560	390	170	93	57	38	26	19						
	2100	520	225	125	77	51	35	24						
	3550	880	380	210	130	86	59	41						

Table 19. Boundaries of Armitage sequential trial with qualitative responses.
(Two tailed, $\alpha = 0.05$; $\beta = 0.05$.)

$\theta_1 = 0.75$							
Unequal pairs n	Boundaries		False positive rate	Middle ends			
	U	L		M'		M''	
				n	y	n	y
9	9	-9	0.004	44	0	44	0
12	10	-10	0.008	62	18	62	-18
15	11	-11	0.012				
18	12	-12	0.015				
20	12	-12	0.020				
23	13	-13	0.023				
26	14	-14	0.026				
28	14	-14	0.029				
31	15	-15	0.031				
34	16	-16	0.033				
37	17	-17	0.034				
39	17	-17	0.036				
42	18	-18	0.037				
45	19	-19	0.038				
47	19	-19	0.039				
50	20	-20	0.039				
53	21	-21	0.040				
56	22	-22	0.040				
58	22	-22	0.041				
60	22	-22	0.042				
61	21	-21	0.044				
62	20	-20	0.047				
$\theta_1 = 0.80$							
8	8	-8	0.008	26	0	26	0
11	9	-9	0.016	40	14	40	-14
14	10	-10	0.022				
17	11	-11	0.027				
20	12	-12	0.031				
23	13	-13	0.033				
26	14	-14	0.035				
29	15	-15	0.037				
32	16	-16	0.038				
35	17	-17	0.039				
38	18	-18	0.040				
39	17	-17	0.042				
40	16	-16	0.047				

Table 19. (Continued)

$\theta_1 = 0.85$							
Unequal pairs	Boundaries		False positive rate	Middle ends			
				M'		M''	
	n	U		L	n	y	n
7	7	-7	0.016	16	0	16	0
11	9	-9	0.022	27	11	27	-11
14	10	-10	0.028				
17	11	-11	0.033				
20	12	-12	0.037				
24	14	-14	0.038				
26	14	-14	0.041				
27	13	-13	0.047				
$\theta_1 = 0.90$							
7	7	-7	0.016	10	0	10	0
10	8	-8	0.029	19	9	19	-9
14	10	-10	0.034				
18	12	-12	0.037				
19	11	-11	0.041				
$\theta_1 = 0.95$							
6	6	-6	0.032	6	0	6	0
11	9	-9	0.038	13	7	13	-7
13	9	-9	0.048				

Table 20. Parameters of Schneiderman-Armitage sequential trial with quantitative responses.

(Two tailed, $\alpha = 0.05$; $\beta = 0.05$)

σ_d	Boundary coefficients		Average "unequal pairs" needed			Average "unequal pairs" with given sample
	c_1	b	\bar{n}_0	\bar{n}_σ	\bar{n}_{\max}	
0.2	18.19	0.10	205	165	270	325
0.3	12.13	0.15	91	74	120	145
0.4	9.09	0.20	51	42	68	82
0.5	7.28	0.25	33	26	43	52
0.6	6.06	0.30	23	18	30	37
0.7	5.20	0.35	17	14	22	27
0.8	4.55	0.40	13	10	17	21
0.9	4.04	0.45	10	8	13	17
1.0	3.64	0.50	8	7	11	13
1.2	3.03	0.60	6	5	8	10
1.4	2.60	0.70	4	3	6	9

Table 21. Coordinates of middle ends of right boundary.*
(Two tailed, $\alpha = \beta = 0.05$, $\sigma_d^2 = 1$, $\delta = 1$)

n'	7.47	8.00	9.00	10.00	11.00	12.00	13.00
y'	0	0.80	1.50	2.10	2.80	3.50	4.30
n'	14.00	15.00	16.00	17.00	18.00	18.50	18.91
y'	5.10	6.00	7.00	8.20	9.70	10.80	13.90

*Note: For any other kinds of sequential trial, the horizontal coordinates n' multiplied with δ^{-2} and the vertical coordinates y' multiplied with σ_d/δ .

Table 22. The basic designs by Latin square.

3 × 3			4 × 4				5 × 5				
A	B	C	A	B	C	D	A	B	C	D	E
B	C	A	B	C	D	A	B	C	D	E	A
C	A	B	C	D	A	B	C	D	E	A	B
			D	A	B	C	D	E	A	B	C
							E	A	B	C	D

6 × 6						7 × 7						
A	B	C	D	E	F	A	B	C	D	E	F	G
B	C	D	E	F	A	B	C	D	E	F	G	A
C	D	E	F	A	B	C	D	E	F	G	A	B
D	E	F	A	B	C	D	E	F	G	A	B	C
E	F	A	B	C	D	E	F	G	A	B	C	D
F	A	B	C	D	E	F	G	A	B	C	D	E
						G	A	B	C	D	E	F

Table 23. Random numbers.

No.	1—10	11—20	21—30	31—40	41—50
1	22 17 68 65 81	68 95 23 92 35	87 02 22 57 51	61 09 43 95 06	58 24 82 03 47
2	19 36 27 59 46	13 79 93 37 55	39 77 32 77 09	85 52 05 30 62	47 83 51 62 74
3	16 77 23 02 77	09 61 87 25 21	28 06 24 25 93	16 71 13 59 78	23 05 47 47 25
4	78 43 76 71 61	20 44 90 32 64	97 67 63 99 61	46 38 03 93 22	69 81 21 99 21
5	03 28 28 26 08	73 37 32 04 05	69 30 16 09 05	88 69 58 28 99	35 07 44 75 47
6	93 22 53 64 39	07 10 63 76 35	87 03 04 79 88	08 13 13 85 51	55 34 57 72 69
7	78 76 58 54 74	92 38 70 96 92	52 06 79 79 45	82 63 18 27 44	69 66 92 19 09
8	23 68 35 26 00	99 53 93 61 28	52 70 05 48 34	56 65 05 61 86	90 92 10 70 80
9	15 39 25 70 99	93 86 52 77 65	15 33 59 05 28	22 87 26 07 47	86 96 98 29 06
10	58 71 96 30 24	18 46 23 34 27	85 13 99 24 44	49 18 09 79 49	74 16 32 23 02
11	57 35 27 33 72	24 53 63 94 09	41 10 76 47 91	44 04 95 49 66	39 60 04 59 81
12	48 50 86 54 48	22 06 34 72 52	82 21 15 65 20	33 29 64 71 11	15 91 29 12 03
13	61 96 48 95 03	07 16 39 33 66	98 56 10 56 79	77 21 30 27 12	90 49 22 23 62
14	36 93 89 41 26	29 70 83 63 51	99 74 20 52 36	87 09 41 15 09	98 60 16 03 03
15	18 87 00 42 31	57 90 12 02 07	23 47 37 17 31	54 08 01 88 63	39 41 88 92 10
16	88 56 53 27 59	33 35 72 67 47	77 34 55 45 70	08 18 27 38 90	16 95 86 70 75
17	09 72 95 84 29	49 41 31 06 70	42 38 06 45 18	64 84 73 31 65	52 53 37 97 15
18	12 96 88 17 31	65 19 69 02 83	60 75 86 90 68	24 64 19 35 51	56 61 87 39 12
19	85 94 57 24 16	92 09 84 38 76	22 00 27 69 85	29 81 94 78 70	21 94 47 90 12
20	38 64 43 59 98	98 77 87 68 07	91 51 67 62 44	40 98 05 93 78	23 32 65 41 18

Table 23. (Continued)

No.	1-10	11-20	21-30	31-40	41-50
21	53 44 09 42 72	00 41 86 79 79	68 47 22 00 20	35 55 31 51 51	00 83 63 22 55
22	40 76 66 26 84	57 99 99 90 37	36 63 32 08 58	37 40 13 68 97	87 64 81 07 83
23	02 17 79 18 05	12 59 52 57 02	22 07 90 47 03	28 14 11 30 79	20 69 22 40 98
24	95 17 82 06 53	31 51 10 96 46	92 06 88 07 77	56 11 50 81 69	40 23 72 51 39
25	35 76 22 42 92	96 11 83 44 80	34 68 35 48 77	33 42 40 90 60	73 96 53 97 86
26	26 29 13 56 41	85 47 04 66 08	34 72 57 59 13	82 43 80 46 15	38 26 61 70 04
27	77 80 20 75 82	72 82 32 99 90	63 95 73 76 63	89 73 44 99 05	48 67 26 43 18
28	46 40 66 44 52	91 36 74 43 53	30 82 13 54 00	78 45 63 98 35	55 03 36 67 68
29	37 56 08 18 09	77 53 84 46 47	31 91 18 95 58	24 16 74 11 53	44 10 13 85 57
30	61 65 61 68 66	37 27 47 39 19	84 83 70 07 48	53 21 40 06 71	95 06 79 88 54
31	93 43 69 64 07	34 18 04 52 35	56 27 09 24 86	61 85 53 83 45	19 90 70 99 00
32	21 96 60 12 99	11 20 99 45 18	48 13 93 55 34	18 37 79 49 90	65 97 38 20 46
33	95 20 47 97 97	27 37 83 28 71	00 06 41 41 74	45 89 09 39 84	51 67 11 52 49
34	97 86 21 78 73	10 65 81 92 59	58 76 17 14 97	04 76 62 16 17	17 95 70 45 80
35	69 92 06 34 13	59 71 74 17 32	27 55 10 24 19	23 71 82 13 74	63 52 52 01 41

Table 23. (Continued)

No.	1-10	11-20	21-30	31-40	41-50
36	04 31 17 21 56	33 73 99 19 87	26 72 39 27 67	53 77 57 68 93	60 61 97 22 61
37	61 06 98 03 91	87 14 77 43 96	43 00 65 68 50	45 60 33 01 07	98 99 46 50 47
38	85 93 85 86 88	72 87 08 62 40	16 06 10 89 20	23 21 34 74 97	76 38 03 29 63
39	21 74 32 47 45	73 96 07 94 52	09 65 90 77 47	25 76 16 19 33	53 05 70 53 30
40	15 69 53 82 80	79 96 23 53 10	65 39 07 16 29	45 33 02 43 70	02 87 40 41 45
41	02 89 08 04 49	20 21 14 68 86	87 63 93 95 17	11 29 01 95 80	35 14 97 35 33
42	87 18 15 89 79	85 43 01 72 73	08 61 74 51 69	89 74 39 82 15	94 51 33 41 67
43	98 83 71 94 22	59 97 50 99 52	08 52 85 08 40	87 80 61 65 31	91 51 80 32 44
44	10 08 58 21 66	72 68 49 29 31	89 85 84 46 06	59 73 19 85 23	65 09 29 75 63
45	47 90 56 10 08	88 02 84 27 83	42 29 72 23 19	66 56 45 65 79	20 71 53 20 25
46	22 85 61 68 90	49 64 92 85 44	16 40 12 89 88	50 14 49 81 06	01 82 77 45 12
47	67 80 43 79 33	12 83 11 41 16	25 58 19 68 70	77 02 54 00 52	53 43 37 15 26
48	27 62 50 96 72	79 44 61 40 15	14 53 40 65 39	27 31 58 50 28	11 39 03 34 25
49	33 78 80 87 15	38 30 06 38 21	14 47 47 07 26	54 96 87 53 32	40 36 40 96 76
50	13 13 92 66 99	47 24 49 57 74	32 25 43 62 17	10 97 11 69 84	99 63 22 32 98

Appendix III

Datasets of Some Real Medical Examples

In this appendix, 12 examples from real world are presented, of which some have been mentioned in the text or exercises. The filename of dataset, the definition and format of variables will be given after each of the examples.

Example 1 A chemical laboratory measured a biochemical index every day as listed the following table. Draw a curve with the following expression:

$$Y = p_1 * \exp[-\exp(p_2 - p_3 * TIME)]$$

Time (day)	Y (biochemical index)	Time (day)	Y (biochemical index)
1	16.080	9	590.030
2	33.830	10	651.920
3	65.800	11	724.930
4	97.200	12	699.560
5	191.550	13	689.960
6	326.200	14	637.560
7	386.870	15	717.410
8	520.53		

Example 2 In order to repair ear injury, a department of orthopaedics measured five variables of the uninjured ear for 300 patients who had ear injury. The five variables were: ear's length (EC), ear's width (EK), the abduction distance of the ear (EZ), the type of ear (EX) (coded from 1 to 6) and the type of auricular lobule (ECX) (coded from 1 to 4). Two other indices were calculated as follows:

$$\text{The ear index (EI)} = EK/EC \times 100\%,$$

$$\text{The abduction index (AI)} = EZ/EK \times 100\%.$$

A cluster analysis had been carried out, according to the five variables: EC, EK, EZ, EI, AI to seek for standard types of the ear for clinical repair of ear. If only one side of ear was injured, the type of standardized ear could be selected according to the normal side. If both ears were injured, the type of standardized ear could be selected according to the judgment of doctors.

The last column of data was the type of ear (TYPE), which was the result of cluster analysis.

The original data were listed as follows:

No.	EC	EK	EZ	EX	ECX	TYPE
1	6.6	3.5	1.9	5	3	1
2	5.9	3.0	2.1	2	2	1
...
300	6.5	3.2	1.5	5	2	4

The dataset, DATA2.DAT, can be found in the appendices at the website.

Example 3 A department of orthodontics studied the diagnostic classification for anterior crossbite in early permanent dentition. The calibration data included 50 patients with anterior crossbite in early permanent dentition. A total of 25 indices about craniofacial construe characters of anterior crossbite were calculated based on X-ray of central occlusion (for example: ARGO, SN, ... , RANG, CV value). The diagnosis of patients (named as TYPE and TYPE 2) was carried out by specialist who had at least three years of clinical experience in orthodontics.

TYPE	1 = skeletal pattern	TYPE 2	1 = maxillary pattern
	2 = denture pattern		2 = mandibular pattern
	0 = similar skeletal pattern		0 = not certain pattern

TYPE 2 was re-classification for skeletal pattern and similar skeletal pattern.

Carry out a stepwise linear discriminant analysis for the classification of TYPE according to the 25 indices mentioned above.

The dataset, DATA3.DAT, can be found in the appendices at the website.

Example 4 A department of internal medicine conducted a clinical research to investigate the relationship between coronary heart disease,

angina pectoris, myocardial infarction and the function of heart. The indices describing the function of heart were indicated by LPA, TC, TG, LDL, HDL, APOA and APOB. The statue of coronary heart disease, angina pectoris and myocardial infarction were indicated by the variables CHD, AP and MI respectively ("no" was coded as 0 and "yes" was coded as 1). The data also included demographic variables such as gender (SEX, female was coded as 0 and male was coded as 1) and age. The total number of observed subjects was 136.

The dataset, DATA4.DAT, can be found in the appendices at the website.

Example 5 In order to investigate the effect of the intraepithelial hyperplasia of esophagus on carcinoma, a cytological examination was carried out for the residents over 30 years old of a certain community. At the same time, the family history about the death on account of esophagus carcinoma was registered. The period of observation was nine and a half years; the data on the incidence of esophagus carcinoma of those residents were collected and listed as follows:

No.	Age	X4	X5	X6	<i>n</i>	<i>r</i>
1	1	0	0	0	11493	62
2	1	0	0	1	44	1
3	1	1	0	0	17	0
4	1	1	0	1	1	1
5	1	0	1	0	900	10
6	1	0	1	1	7	0
7	2	0	0	0	7128	195
8	2	0	0	1	30	2
9	2	1	0	0	78	3
10	2	1	0	1	0	0
11	2	0	1	0	417	17
12	2	0	1	1	5	0
13	3	0	0	0	1115	34
14	3	0	0	1	7	0
15	3	1	0	0	18	3
16	3	1	0	1	1	0
17	3	0	1	0	27	2
18	3	0	1	1	2	1

The number of observed subjects was N , and the number of patients was R . The code sheet for this dataset was:

Age	1 = 30–49 years old 2 = 50–69 years old 3 = 70 years old and above	Family history (X5)	1 = spouse died from carcinoma of esophagus 0 = others
Family history (X4)	1 = relative died from carcinoma of esophagus 0 = others	Severe epithelial hyperplasia of esophagus (X6)	1 = severe hyperplasia 0 = no

Example 6 The esophagus carcinoma, stomach carcinoma, colon and rectal carcinoma were most common all over the world among the cancers of digestive tract. Recently, some new diagnosis technique such as fiberoptic endoscope, X-ray double contrast neoplasty and exfoliative cytological examination made it possible to early diagnose for these cancers, but the incidence and prognosis had not been changed significantly. To investigate the prognosis situation of the patients with stomach carcinoma, data were collected by an epidemiological survey (total 107 patients, 75 deaths and 32 survivals). The code sheet for the variables to be considered was given as follows:

Variable	Code	Variable	Code
Gender	0 = male, 1 = female	Age	1 = 28–40, 2 = 41–50, 3 = 51–60, 4 = 61–
Method of operation (OM)	1 = radical gastrectomy 2 = palliative gastrectomy	Location of cancer (SITE)	1 = gastric antrum 2 = body of stomach 3 = cardia of stomach
Type based on gross inspection (gross)	1 = type I, 2 = type II 3 = type III, 4 = type IV	Degree of differentiation (DD)	0 = high differentiation 1 = lower differentiation

Variable	Code	Variable	Code
Method of growth (GM)	0 = expansive growth 1 = infiltrative growth	Depth of infiltration (DI)	1 = muscle layer, 2 = serous coat layer, 3 = out layer of serous coat
Lymphatic metastasis (LN)	0 = none, 1 = one lymph node 2 = two lymph node	Lymphatic infiltration (LI)	0 = none, 1 = have
Distance metastasis (DM)	0 = none, 1 = have	LC and PC reaction of infiltration (IL)	0 = none, 1 = mild, 2 = obvious
DNA (DNA)	0 = diploid, 1 = allosome	Stage of TNM (TNM)	1 = Ib, 2 = II, 3 = IIIa, 4 = IIIb, 5 = IV
Outcome (SURV)	0 = death, 1 = survival	Survival month (LIFE)	

The dataset, DATA6.DAT, can be found in the appendices at the website.

Example 7 To compare the efficacy of Salmeterol (group C), Salbutamol (group B) and placebo (group A), a department of pulmonary randomly assigned the patients into three treatment groups and measured the PEFR value at different time under the condition of double-blind. The results were listed in the following table. Test whether there was a significant difference among the three treatments.

Case no.	Group A, B or C	Measurement at night				Measurement in day time			
		Before NP	1 week NT1	2 week NT2	3 week NF	Before DP	1 week DT1	2 week DT2	3 week DF
1									
2									
...									

The dataset, DATA7.DAT, can be found in the appendices at the website.

Example 8 To investigate the relationship between blood pressure, cholesterol serum level and coronary heart disease, 1330 patients were cross-classified into a $2 \times 4 \times 4$ contingency table according to the following criteria. Analyze the data by the methods for discrete variables.

Coronary heart disease	Blood pressure (mm Hg)		Cholesterol serum level			
			<200	200–219	220–259	≥260
yes	1	<127	2	3	3	4
	2	127–146	3	2	1	3
	3	147–166	8	11	6	6
	4	≥167	7	12	11	11
no	1	<127	117	121	47	22
	2	127–146	85	98	43	20
	3	147–166	119	209	68	43
	4	≥167	67	99	46	33

Example 9 In order to investigate the efficacy of tretinoin for leukemia, there was a clinical follow-up study including three drug groups. The grouping variable was GRO (simple chemotherapy group coded 1, tretinoin group coded 2 and chemotherapy plus tretinoin group coded 3). The outcome was DEAD (death coded 1), and the variable name for follow up time was LIFE (month).

The dataset, DATA9.DAT, can be found in the appendices at the website.

Example 10 The following table showed the measurement results of sex hormone at several time points for a woman. Fit a curve to express the sex hormone changing with time. The expression for the curve is assumed to be:

$$Y = p_1 \exp(p_2 X) + p_3 \exp(p_4 X),$$

where p_1, p_2, p_3, p_4 are the parameters to be estimated.

X (hour)	y	X (hour)	y
0	641	120	236
1	866	180	226
5	891	240	224
10	947	300	130
15	757	360	128
20	735	480	117
25	541	720	143
30	476	1440	85
40	590	2880	84
50	369	4320	152
60	352	7200	139
90	295		

Example 11 In order to evaluate the severity of retina disease, a department of ophthalmology measured 10 variables in 131 diabetes patients, including age (AGE), course of diabetes (TIME), glucose level (GLUCOSE), visual acuity (VISION) and A-wave latency (AT), A-wave value (AV), B-wave latency (BT), B-wave value (BV), QP-wave latency (QPT), QP-wave value (QPV) of electroretinogram. At the same time, the details of the retinal changes were also examined. According to the unified criteria, the severity of retina disease was diagnosed as mild, moderate or severe (GROUP, coded as A1, A2 and A3 respectively). The raw data were saved in EYE1.DAT as calibration data. In addition, the above variables of another 31 diabetic patients were measured and the data were saved in EYE2.DAT as test data.

Find a discriminant function by stepwise discriminant analysis based on the training sample EYE1.DAT, and classify the test data EYE2.DAT using the linear discriminant function.

Both EYE1.DAT and EYE2.DAT were free format and can be found in the appendices at the website.

Example 12 In order to classify and identify bacterium, a research constitute analyzed the content of fatty acids in bacterium cell using gas chromatography. 24 bacteria were collected, including eight jejuno-campylo bacterium (named CJ1–CJ8), three colonic campylo bacterium (named CC1–CC3), nine pyloro-spirillosis (named HP1–HP9) and four other enteric bacilli (named XX1–XX4). The percentile contents of 12 different fatty acids (named X1–X12) of those bacteria were measured. The data were saved in BACTERIA.DAT (can be found in the appendices at the website).

Classify the 24 bacteria by cluster analysis based on the 12 variables.

(1st edn. Binghua Su, Qingbo He; 2nd edn. Jing Gu)





