

С. А. Айвазян
В. С. Мхитарян

ТЕОРИЯ ВЕРОЯТНОСТЕЙ и ПРИКЛАДНАЯ СТАТИСТИКА

Том **1**

*Рекомендовано Министерством общего
и профессионального образования Российской
Федерации в качестве **учебника** для студентов
экономических специальностей высших
учебных заведений*



Москва
2001

УДК 311+330.43(075.8)

ББК 60.6+65в6я73

П85

Первое издание данного учебника подготовлено в рамках Проекта Tacis "Преподавание экономических и бизнес-дисциплин в средних школах, технических и классических университетах", реализованного Фондом экономических исследований Университета "Эразмус", Роттердам (SEOR/EUR) и Государственным университетом — Высшей школой экономики

Рецензенты:

*кафедра статистики С.-Петербургского государственного университета
экономики и финансов; проф. Э.Б. Ершов (ГУ ВШЭ) и проф. Я. Магнус (Университет Тилбурга)*

Главный редактор издательства *Н.Д. Эриашвили*

П85 **Прикладная статистика. Основы эконометрики: Учебник для вузов: В 2 т.**
2-е изд., испр. — Т. 1: *Айвазян С.А., Мхитарян В.С.* Теория вероятностей
и прикладная статистика. — М.: ЮНИТИ-ДАНА, 2001. — 656 с.
ISBN 5-238-00304-8

Содержание и стиль изложения в учебнике соответствуют принятым Министерством образования РФ стандартам и учебным программам высших учебных заведений экономического профиля по дисциплинам "Теория вероятностей", "Математическая статистика" и "Многомерные статистические методы" (или "Многомерный статистический анализ"). При этом первые две дисциплины входят в учебные планы 1-й ступени образования (бакалавриата), а третья может присутствовать (в зависимости от конкретного вуза) в учебных планах бакалавриата или магистратуры. Усвоение включенного в этот том материала предусматривает для каждой из упомянутых дисциплин общий объем аудиторных занятий, равный приблизительно 64 часам (32 часа лекций и 32 часа практических занятий). Изложение построено таким образом, чтобы добиться цельного (системного) восприятия всего блока эконометрических дисциплин, представленных в двух томах второго издания: упомянутые три дисциплины первого тома дополнены во втором томе широким набором моделей регрессионного анализа, методами и моделями анализа временных рядов и методами построения и анализа систем одновременных уравнений.

Для студентов, аспирантов, преподавателей, а также специалистов по прикладной статистике и эконометрике.

ББК 60.6+65в6я73

Учебник

Айвазян Сергей Артемьевич, Мхитарян Владимир Сергеевич

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И ПРИКЛАДНАЯ СТАТИСТИКА

Оформление художника *А.В. Лебедева*

Лицензия серия ИД № 03562 от 19.12.2000 г. Подписано в печать 10.08.2001. Формат 70х100 1/16
Усл. печ. л. 53,3. Уч.-изд. л. 43. Тираж 20 000 экз. (1-й завод — 5000). Заказ 1639

ООО «ИЗДАТЕЛЬСТВО ЮНИТИ-ДАНА». Генеральный директор *В.Н. Закаидзе*

123298, Москва, ул. Ирины Левченко, 1—9. Тел. (095) 194-00-15. Тел./факс (095) 194-00-14
www.unity-dana.ru E-mail: unity@msm.ru

Отпечатано во ФГУП ИПК «Ульяновский Дом печати». 432980, г. Ульяновск, ул. Гончарова, 14
Качество печати соответствует предоставленным оригиналам

ISBN 5-238-00304-8

© С.А. Айвазян, В.С. Мхитарян, 1998, 2001
© «ИЗДАТЕЛЬСТВО ЮНИТИ-ДАНА», 1998, 2001
Воспроизведение всей книги или любой ее
части запрещается без письменного
разрешения издательства

ПРИКЛАДНАЯ СТАТИСТИКА

ОСНОВЫ ЭКОНОМЕТРИКИ

В двух томах

ИЗДАНИЕ ВТОРОЕ, ИСПРАВЛЕННОЕ

S. A. Aivazian
V. S. Mkhitarian

PROBABILITY THEORY AND APPLIED STATISTICS

Volume **1**

Textbook



Moscow
2001

О Г Л А В Л Е Н И Е

К читателю	15
Предисловие к первому изданию	16
Предисловие ко второму изданию	20
В в е д е н и е. Вероятностно-статистические методы в моделировании социально-экономических процессов и анализе данных .	23
В.1. Математико-статистический инструментарий экономических исследований	24
В.1.1. Назначение и составные части учебника	24
В.1.2. Прикладная статистика	26
В.1.3. Теория вероятностей и математическая статистика . .	28
В.2. Теоретико-вероятностный способ рассуждения в прикладной статистике и эконометрике	29
В.2.1. Границы применимости теоретико-вероятностного способа рассуждения	29
В.2.2. Что дает объединение теоретико-вероятностного и статистического способов рассуждения?	35
В.3. Вероятностно-статистическая (эконометрическая) модель как частный случай математической модели	40
В.3.1. Математическая модель	40
В.3.2. Основные этапы вероятностно-статистического моделирования	43
В.3.3. Моделирование механизма явления вместо формальной статистической фотографии	45
Выводы	48

Раздел I: Основы теории вероятностей	51
Глава 1. Правила действий со случайными событиями и вероятностями их осуществления	52
1.1. Дискретное вероятностное пространство	52
1.1.1. Процесс регистрации наблюдения на объекте исследуемой совокупности (случайный эксперимент)	52
1.1.2. Случайные события и правила действий с ними	53
1.1.3. Вероятностное пространство. Вероятности и правила действий с ними	58
1.2. Непрерывное вероятностное пространство (аксиоматика А. Н. Колмогорова)	69
1.2.1. Специфика общего (непрерывного) случая вероятностного пространства	69
1.2.2. Случайные события, их вероятности и правила действий с ними (аксиоматический подход А. Н. Колмогорова)	71
Выводы	75
Глава 2. Случайные величины (исследуемые признаки)	77
2.1. Определение и примеры случайных величин	77
2.2. Возможные и наблюдаемые значения случайной величины	79
2.3. Типы случайных величин	80
2.4. Одномерные и многомерные (совместные) законы распределения вероятностей случайных величин	83
2.5. Способы задания закона распределения: функция распределения, функция плотности	89
2.5.1. Функция распределения вероятностей одномерной случайной величины	89
2.5.2. Функция плотности вероятности одномерной случайной величины	92
2.5.3. Многомерные функции распределения и плотности. Статистическая независимость случайных величин	94
2.6. Основные числовые характеристики случайных величин	98
2.6.1. Понятие о математических ожиданиях и моментах	99
2.6.2. Характеристики центра группирования значений случайной величины	102
2.6.3. Характеристики степени рассеяния значений случайной величины	104
2.6.4. Квантили и процентные точки распределения	106
2.6.5. Асимметрия и эксцесс	108

2.6.6. Основные характеристики многомерных распределений (ковариации, корреляции, обобщенная дисперсия и др.) .	109
Выводы	112
Глава 3. Модели законов распределения вероятностей, наиболее распространенные в практике статистических исследований	114
3.1. Законы распределения, используемые для описания механизмов генерации реальных социально-экономических данных	115
3.1.1. Распределения, возникающие при анализе последовательности испытаний Бернулли: биномиальное и отрицательное биномиальное	115
3.1.2. Гипергеометрическое распределение	119
3.1.3. Распределение Пуассона	121
3.1.4. Полиномиальное (мультиномиальное) распределение	123
3.1.5. Нормальное (гауссовское) распределение	125
3.1.6. Логарифмически-нормальное распределение	129
3.1.7. Равномерное (прямоугольное) распределение	132
3.1.8. Распределения Вейбулла и экспоненциальное (показательное)	134
3.1.9. Распределение Парето	139
3.1.10. Распределение Коши	140
3.2. Законы распределения вероятностей, используемые при реализации техники статистических вычислений	141
3.2.1. «Хи-квадрат»-распределение с m степенями свободы ($\chi^2(m)$ -распределение)	142
3.2.2. Распределение Стьюдента с m степенями свободы ($t(m)$ - распределение)	143
3.2.3. Распределение дисперсионного отношения с числом степеней свободы числителя m_1 и числом степеней свободы знаменателя m_2 ($F(m_1, m_2)$ - распределение)	145
3.2.4. Гамма-распределение (Г-распределение)	147
3.2.5. Бета-распределение (β -распределение)	148
Выводы	151
Глава 4. Основные результаты теории вероятностей	153
4.1. Неравенство Чебышева	153
4.2. Закон больших чисел и его следствия	155
4.2.1. Закон больших чисел	155
4.2.2. Теорема Бернулли	156

4.3. Особая роль нормального распределения: центральная предельная теорема	157
4.3.1. Центральная предельная теорема	158
4.3.2. Многомерная центральная предельная теорема	159
4.3.3. Комментарии к центральной предельной теореме	159
4.4. Законы распределения вероятностей случайных признаков, являющихся функциями от известных случайных величин	161
Выводы	166
Глава 5. Цепи Маркова	168
5.1. Последовательности случайных экспериментов и случайных величин в дискретном вероятностном пространстве	168
5.2. Последовательности, образующие цепь Маркова (определения, примеры, прикладные задачи)	170
5.3. Основные характеристики и свойства цепей Маркова	177
5.3.1. Основные характеристики	177
5.3.2. Классификация состояний и цепей	179
5.3.3. Свойства цепей Маркова	182
5.4. Анализ некоторых задач и примеров	185
Выводы	190
Раздел II: Основы математической статистики	193
Глава 6. Основы статистического описания и статистика нормального закона	194
6.1. Генеральная совокупность, выборка из нее и основные способы организации выборки	194
6.2. Основные выборочные характеристики и их свойства	200
6.2.1. Выборочные (эмпирические) функции распределения, относительные частоты и функции плотности	201
6.2.2. Выборочные аналоги начальных и центральных моментов случайной величины	207
6.2.3. Эмпирические аналоги центра группирования генеральной совокупности	208
6.2.4. Эмпирические аналоги показателей вариации рассеивания случайной величины	209
6.2.5. Выборочные коэффициенты асимметрии и эксцесса	210
6.2.6. Статистическая устойчивость выборочных характеристик	214
6.2.7. Асимптотически-нормальный характер случайного варьирования основных выборочных характеристик	216

6.2.8. Поведение выборочных характеристик в нормальной генеральной совокупности (статистика нормального закона)	219
6.3. Вариационный ряд и порядковые статистики	224
6.3.1. Закон распределения вероятностей i -го члена вариационного ряда	225
6.3.2. Совместные (многомерные) распределения членов вариационного ряда	227
6.3.3. Порядковые статистики как эмпирические (выборочные) аналоги квантилей и процентных точек распределения	229
Выводы	229
Глава 7. Статистическое оценивание параметров	231
7.1. Начальные сведения о задаче статистического оценивания параметров	232
7.1.1. Постановка задачи	232
7.1.2. Статистики, статистические оценки, их основные свойства	233
7.1.3. Состоятельность	234
7.1.4. Несмещенность	236
7.1.5. Эффективность	238
7.2. Функция правдоподобия. Количество информации, содержащееся в n независимых наблюдениях относительно неизвестного значения параметра	241
7.3. Неравенство Рао–Крамера–Фреше и измерение эффективности оценок	244
7.4. Понятие об интервальном оценивании и доверительных областях (постановка задач)	248
7.5. Методы статистического оценивания неизвестных параметров	249
7.5.1. Метод максимального (наибольшего) правдоподобия	249
7.5.2. Метод моментов	258
7.5.3. Оценивание с помощью «взвешенных» статистик; цензурирование, урезание выборок и порядковые статистики как частный случай взвешивания	261
7.5.4. Построение интервальных оценок (доверительных областей)	263
7.6. Байесовский подход к статистическому оцениванию	269
7.6.1. «Философия» байесовского подхода	269
7.6.2. Общая логическая схема и базовые формулы байесовского метода оценивания параметров	270
7.6.3. Примеры байесовского оценивания	273
Выводы	279
Приложение к гл. 7 (доказательство неравенства информации)	281

Глава 8. Статистическая проверка гипотез (статистические критерии)	283
8.1. Основные типы гипотез, проверяемых в ходе статистического анализа и моделирования	284
8.1.1. Гипотезы о типе закона распределения исследуемой случайной величины	284
8.1.2. Гипотезы об однородности двух или нескольких обрабатываемых выборок или некоторых характеристик анализируемых совокупностей	285
8.1.3. Гипотезы о числовых значениях параметров исследуемой генеральной совокупности	285
8.1.4. Гипотезы об общем виде модели описывающей, статистическую зависимость между признаками	286
8.2. Общая логическая схема статистического критерия	287
8.3. Построение статистического критерия; принцип отношения правдоподобия	290
8.3.1. Сущность принципа отношения правдоподобия	290
8.3.2. Критерий логарифма отношения правдоподобия для проверки простой гипотезы	292
8.3.3. Критерий отношения правдоподобия для проверки сложной гипотезы	293
8.4. Характеристики качества статистического критерия	294
8.5. Последовательная схема принятия решения (последовательные критерии)	297
8.5.1. Последовательная схема наблюдений	297
8.5.2. Последовательный критерий отношения правдоподобия (критерий Вальда) и его свойства	299
8.6. Методы проверки статистических гипотез: примеры статистических критериев	300
8.6.1. Критерии согласия	300
8.6.2. Критерии однородности	309
8.6.3. Проверка гипотез о числовых значениях параметров	317
Выводы	325
 Раздел III: Методы прикладной статистики	 327
Глава 9. Введение в прикладной статистический анализ	328
9.1. Назначение и содержание прикладной статистики	328
9.1.1. Два подхода к интерпретации и анализу исходных статистических данных	328

9.1.2. Три центральные проблемы прикладной статистики . . .	331
9.1.3. Новые постановки задач и ослабление ограничительных условий в канонических математико-статистических и эконометрических моделях	338
9.2. Основные этапы прикладного статистического анализа . . .	339
Выводы	346
 Г л а в а 10. Статистическое исследование зависимостей (основ- ные понятия и постановки задач)	 349
10.1. Общая формулировка проблемы, пример	349
10.2. Какова конечная прикладная цель статистического исследова- ния зависимостей	359
10.3. Математический инструментарий	362
10.4. Некоторые типовые задачи практики эконометрического моде- лирования	364
10.5. Основные типы зависимостей между количественными пере- менными	370
10.6. Основные этапы статистического исследования зависимостей .	375
10.7. Выбор общего вида функции регрессии	382
10.7.1. Использование априорной информации о содержатель- ной сущности анализируемой зависимости	383
10.7.2. Предварительный анализ геометрической структуры исходных данных	385
10.7.3. Статистические критерии проверки гипотез об общем виде функции регрессии	386
10.7.4. Некоторые общие рекомендации	390
Выводы	392
 Г л а в а 11. Корреляционный анализ многомерной генеральной совокупности	 396
11.1. Назначение и место корреляционного анализа в статистиче- ском исследовании	396
11.2. Корреляционный анализ количественных признаков	398
11.2.1. Коэффициент детерминации как универсальная харак- теристика степени тесноты статистической связи . . .	399
11.2.2. Исследование линейной зависимости y от единственной объясняющей переменной x : парный коэффициент кор- реляции	404
11.2.3. Исследование парных нелинейных связей: корреляцион- ное отношение	412

11.2.4. Исследование линейной зависимости y от нескольких объясняющих переменных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$: множественный и частные коэффициенты корреляции	417
11.3. Корреляционный анализ порядковых (ординальных) переменных: ранговая корреляция	428
11.3.1. Исходные статистические данные (таблица или матрица рангов типа «объект-свойство»)	429
11.3.2. Понятие ранговой корреляции	430
11.3.3. Основные задачи статистического анализа связей между ранжировками	431
11.3.4. Ранговый коэффициент корреляции Спирмэна	432
11.3.5. Ранговый коэффициент корреляции Кендалла	434
11.3.6. Обобщенная формула для парного коэффициента корреляции и связь между коэффициентами Спирмэна и Кендалла	438
11.3.7. Статистические свойства выборочных характеристик парной ранговой связи	439
11.3.8. Коэффициент конкордации (согласованности) как измеритель статистической связи между несколькими порядковыми переменными	442
11.3.9. Проверка статистической значимости выборочного значения коэффициента конкордации	444
11.4. Корреляционный анализ категоризованных переменных: таблицы сопряженности	447
11.4.1. Исходные статистические данные (таблицы сопряженности)	447
11.4.2. Основные измерители степени тесноты статистической связи между двумя категоризованными переменными	448
Выводы	453

Глава 12. Распознавание образов и типологизация объектов в социально-экономических исследованиях (методы классификации)	457
12.1. Сущность, типологизация и прикладная направленность задач классификации объектов	457
12.2. Классификация при наличии обучающих выборок (дискриминантный анализ)	471
12.2.1. Класс как генеральная совокупность и базовая идея вероятностно-статистических методов классификации	471
12.2.2. Функции потерь и вероятности неправильной классификации	472

12.2.3. Принципиальное решение общей задачи построения оптимальных (байесовских) процедур классификации	473
12.2.4. Параметрический дискриминантный анализ в случае нормальных классов	476
12.3. Классификация без обучения (параметрический случай): расщепление смесей вероятностных распределений	479
12.3.1. Понятие смеси вероятностных распределений	480
12.3.2. Задача расщепления смесей распределений	486
12.3.3. Общая схема решения задачи автоматической классификации в рамках модели смеси распределений (сведение к схеме дискриминантного анализа)	487
12.4. Классификация без обучения (непараметрический случай): методы кластер-анализа	488
12.4.1. Общая постановка задачи автоматической классификации	488
12.4.2. Расстояния между отдельными объектами и меры близости объектов друг к другу	491
12.4.3. Расстояния между классами объектов	495
12.4.4. Функционалы качества разбиения на классы и экстремальная постановка задачи кластер-анализа	498
12.4.5. Формулировка экстремальных задач разбиения исходного множества объектов на классы при неизвестном числе классов	503
12.4.6. Основные типы задач кластер-анализа и основные типы кластер-процедур	503
12.4.7. Иерархические процедуры	505
12.4.8. Параллельные кластер-процедуры	507
12.4.9. Последовательные кластер-процедуры	512
Выводы	516

Глава 13. Снижение размерности исследуемого многомерного признака и отбор наиболее информативных показателей 520

13.1. Сущность, типологизация и прикладная направленность задач снижения размерности	520
13.2. Метод главных компонент	526
13.2.1. Основные понятия и определения	526
13.2.2. Вычисление главных компонент	529
13.2.3. Основные числовые характеристики главных компонент	531
13.2.4. Геометрическая интерпретация главных компонент	538
13.2.5. Оптимальные свойства главных компонент	541

13.2.6. Статистические свойства выборочных главных компонент, статистическая проверка некоторых гипотез . .	544
13.2.7. Применение свойств выборочных характеристик главных компонент	547
13.3. Факторный анализ	551
13.3.1. Сущность модели факторного анализа	551
13.3.2. Общий вид линейной модели, ее связь с главными компонентами	552
13.3.3. Основные задачи факторного анализа	556
13.3.4. Вопросы идентификации модели факторного анализа .	558
13.3.5. Статистическое исследование модели факторного анализа	559
13.4. Некоторые эвристические методы снижения размерности . .	570
13.4.1. Природа эвристических методов	570
13.4.2. Метод экспериментальной группировки признаков . .	571
13.4.3. Метод корреляционных плеяд	577
13.5. Построение сводного (интегрального) латентного показателя качества (или эффективности функционирования) сложной системы	580
13.5.1. Общая постановка задачи	580
13.5.2. Сводный показатель («выходное качество») и его целевая функция	581
13.5.3. Исходные данные	583
13.5.4. Алгоритмические и вычислительные вопросы построения неизвестной целевой функции	585
13.5.5. Примеры построения интегрального показателя с помощью экспертно-статистического метода	589
13.6. Многомерное шкалирование	592
13.6.1. Постановка задачи метрического многомерного шкалирования	592
13.6.2. Решение задачи метрического многомерного шкалирования	593
13.6.3. Понятие о неметрическом многомерном шкалировании (МШ)	595
Выводы	595
Приложение 1. Таблицы математической статистики	601
Приложение 2. Необходимые сведения из матричной алгебры	619
Литература	642
Алфавитно-предметный указатель	644

К ЧИТАТЕЛЮ

Рекомендую студентам и преподавателям, специализирующимся в области прикладной статистики и эконометрики, 2-е издание учебника «Прикладная статистика и основы эконометрики», том 1: «Теория вероятностей и прикладная статистика». Книга подготовлена выдающимися специалистами, имеющими богатый опыт преподавания излагаемого предмета и ведущими, наряду с этим, серьезную исследовательскую работу в области эконометрического анализа социально-экономических процессов. «Предисловия» авторов данного издания содержат точную, с моей точки зрения, характеристику основного замысла книги и ее места в ряду аналогичных изданий. Книгу выгодно выделяют, по меньшей мере, два ее свойства. В о - п е р в ы х, отбор материала первого тома, стиль изложения, расставленные в ней акценты подчинены общей идее востребованности излагаемых методов и моделей *именно теорией и практикой эконометрики*. В о - в т о р ы х, содержание и структура учебника отражают более широкий (по сравнению с традиционным) взгляд авторов на состав вероятностно-статистического инструментария эконометрики. В частности, авторы включили в состав этого инструментария и *цепи Маркова* (гл. 5), зарекомендовавшие себя полезным инструментом в исследовании динамики различного рода структурных социально-экономических изменений, и различные многомерные статистические методы (*дискриминантный и кластер анализы*, гл. 12, *методы снижения размерности*, гл. 13), органично дополняющие «законно прописанные» в эконометрике методы и модели регрессионного анализа, и, наконец, *байесовский подход* к идентификации моделей (п. 7.6), повышенная востребованность которого обусловлена именно экономической спецификой статистических исследований, *при которой объем исходных данных (число имеющихся наблюдений), как правило, не слишком превышает размерности анализируемых моделей*.

Много лет я поддерживал самые тесные научные контакты с одним из авторов книги — Сергеем Артемьевичем Айвазяном, в первые пятнадцать лет, — в качестве его научного руководителя (вначале — на механико-математическом факультете МГУ им. М.В. Ломоносова, затем — в Математическом институте им. В.А. Стеклова Российской академии наук). Уже в самом начале своего творческого пути он умел удачно соединять глубокие математические исследования с живым интересом к различным областям приложений вероятностно-статистической науки, в том числе и к экономике. Я твердо уверен, что 2-е издание книги будет встречено читателями с таким же энтузиазмом, как и первое.

Академик Юрий Васильевич ПРОХОРОВ,
заведующий отделом теории вероятностей Математического института им. В.А. Стеклова РАН, заведующий кафедрой математической статистики факультета вычислительной математики и кибернетики МГУ им. М.В. Ломоносова.

25 июля 2001 г.

W. Trifunov

ПРЕДИСЛОВИЕ К ПЕРВОМУ ИЗДАНИЮ

Эконометрика — одна из базовых (наряду с микро- и макроэкономикой) дисциплин экономического образования во всем мире. К сожалению, до начала 90-х годов эконометрика по существу не была признана в СССР и России, не включалась в учебные планы подготовки специалистов (студентов, аспирантов) экономического профиля. Объяснение этому найти нетрудно: из трех основных составляющих эконометрики — экономической теории, экономической статистики и математико-статистического инструментария две первые были представлены в нашей стране явно неудовлетворительно. Не было доброкачественной экономической теории, не было системы национальных счетов и необходимого информационного обеспечения эконометрического моделирования.

Теперь ситуация изменилась. Авторам предлагаемого вниманию читателя учебника довелось принять непосредственное участие в процессе «восстановления в своих законных правах» эконометрики: в формировании базовых положений концепции современного экономического образования в российской высшей школе, в составлении первых программ и чтении первых курсов лекций по этой дисциплине на экономическом факультете Московского государственного университета им. М. В. Ломоносова (МГУ), начиная с 1992 г., и в Московском государственном университете экономики, статистики и информатики (МЭСИ), — с 1993 г. В формировании своей позиции по данной проблеме авторы опирались на многолетний опыт исследовательской и педагогической работы в области разработки и практического использования методов эконометрического моделирования: один из них (С. А. Айвазян), работая с 1969 года в Центральном экономико-математическом институте Российской академии наук, занимается эконометрическим моделированием распределительных отношений в обществе, одновременно являясь автором учебных программ и постоянным лектором по всему спектру дисциплин эконометрического профиля — *теории вероятностей, математической статистике, многомерному статистическому анализу* (МГУ им. М. В. Ломоносова, МЭСИ, Российская экономическая школа — РЭШ), *методам прогнозирования в бизнесе* (Московское отделение Калифорнийского государственного университета, г. Хэйвард); другой (В. С. Мхитарян) использует эконо-

метрические методы в социально-экономических исследованиях и задачах статистического контроля качества изделий, а также много лет читает курсы лекций вероятностно-статистического профиля в МЭСИ.

Что побудило авторов написать этот учебник? Ведь богатые традиции европейских и северо-американских университетов в области эконометрической науки и ее преподавания в системе экономического образования всех уровней (undergraduate, graduate, post-graduate) отражены в разнообразной и интересной монографической и учебно-методической литературе по данному предмету. Достаточно назвать широко распространенные учебники Дж. Джонстона, Э. Маленво, А. Goldberger, R. S. Pindyck—D. L. Rubinfeld, W. Greene, C. Dougherty, E. R. Berndt (см. список литературы в конце книги). Почему бы не повторить переводы двадцатилетней давности первых двух учебников и не перевести с английского языка некоторые из других, здесь упомянутых? Мы полагаем, что в условиях острого дефицита на отечественном рынке эконометрической учебной литературы эта деятельность была бы крайне желательной, более того — необходимой. Однако издание нашей книги имеет свой замысел, свою специфику, и вот в чем они заключаются.

Во-первых, в предлагаемом учебнике отражено понимание содержания математико-статистического инструментария эконометрики, несколько отличающееся от общепринятого. По нашему мнению, современные достижения математико-статистической науки (особенно в многомерном статистическом анализе), с одной стороны, и существенное расширение круга экономических задач, требующих эконометрических методов решения, — с другой, обусловили *необходимость более широкого взгляда на математико-статистический инструментарий эконометрики* и, в частности, включения в него, помимо традиционных разделов по регрессионным моделям, анализу временных рядов и системам одновременных уравнений, таких разделов многомерного статистического анализа, как марковские цепи, классификация многомерных наблюдений и снижение размерности анализируемого факторного пространства. Говоря о широком спектре экономических задач, требующих выходящих за традиционные рамки эконометрических методов решения, мы имели в виду, в частности, статистическое исследование динамики структурных изменений (в демографии, в стратификационной структуре общества и т.п.), выявление скрытых (латентных) факторов, определяющих течение того или иного социально-экономического процесса, построение интегральных индикаторов качества или эффективности функционирования социально-экономической системы, типологизацию социально-экономических объектов и др.

Во-вторых, в ходе многолетнего опыта преподавания различных дисциплин вероятностно-статистического профиля в экономических вузах и на экономических факультетах университетов мы пришли к убеждению, что *необходимо строить учебный процесс таким образом, чтобы добиваться цельного, системного восприятия всего блока этих дисциплин*. Речь идет, в частности, о курсах по элементарным методам статистической обработки данных (или дескриптивной статистике), теории вероятностей, математической статистике, многомерному статистическому анализу (или многомерным статистическим методам), анализу временных рядов и, наконец, эконометрике. Очевидно, *реализации этой цели должен способствовать и учебник, одновременно содержащий взаимосвязанное изложение всех этих курсов*.

Другими словами, мы попытались написать такую книгу, которую нам бы хотелось иметь под рукой в процессе нашей преподавательской деятельности. К сожалению, среди многих прекрасных зарубежных книг по эконометрике книги, обладающей двумя вышеуказанными особенностями, не оказалось.

Заметим, что несмотря на наличие ряда иллюстративных примеров и упражнений, предлагаемый учебник *не решает проблемы задачника по эконометрике*. Поэтому для проведения полноценного учебного процесса он должен быть дополнен набором эконометрических задач и упражнений (например, в духе книги [Berndt, E. R.]).

Материал учебника и ответственность распределены между авторами следующим образом. В. С. Мхитарян принимал участие в написании глав 6, 7, 8 и 13, а также предложил большую часть задач, которыми снабжены главы учебника. Остальной материал (включая упомянутые главы) написаны С. А. Айвазяном. Им же выполнено общее научное редактирование учебника.

В заключение мы хотим выразить свою признательность. Мы благодарны, во-первых, руководителям образовательного проекта EU-TACIS «Преподавание экономических и бизнес дисциплин в средних школах, технических и классических университетах» профессору Соломону Кохену (S. Cohen, SEOR и Эразмус Университет, Роттердам, Голландия) и ректору Государственного университета-Высшей школы экономики — профессору Ярославу Ивановичу Кузьминову: финансовая поддержка этого проекта сыграла решающую роль в том, чтобы авторы смогли определить для себя *приоритетной* задачей написание данного учебника. Мы выражаем также свою признательность нашим коллегам: профессору эконометрики Тилбургского университета и Высшей экономической школы Лондона Яну Магнусу и профессору Государственного университета-Высшей

школы экономики — Эмилю Борисовичу Ершову. Их внимание к учебнику было постоянным (в процессе его написания), взыскательным и одновременно доброжелательным. Высказанные ими замечания и советы, бесспорно, способствовали улучшению качества рукописи. Существенное влияние на замысел и содержание книги оказал опыт исследовательской и педагогической работы авторов, их постоянные контакты с коллегами по ЦЭМИ РАН, по научному семинару «Многомерный статистический анализ и вероятностное моделирование реальных процессов». Без них и без наших главных критиков и генераторов вопросов — многих поколений студентов МГУ им. М. В. Ломоносова, МЭСИ, РЭШ, — эта книга вряд ли появилась бы на свет. Мы благодарны Алле Павловне и Галине Юрьевне Грохотовым за самоотверженный труд по подготовке оригинал-макета рукописи книги, а также Николаю Владимировичу Третьякову и Елене Владимировне Герасимовой за очень полезную и профессиональную консультационную поддержку Грохотовых в этом объемном и непростом производственном процессе. Что касается слабых мест и недостатков учебника, то за них всю ответственность несут, естественно, только авторы. Мы будем признательны читателям за их отзывы о книге и критические замечания, направленные в издательство или непосредственно нам.

Москва–Роттердам–
деревня «Плужково»
Московской области.
1996–1998 гг.

*С. А. Айвазян
В. С. Мхитарян*

ПРЕДИСЛОВИЕ КО ВТОРОМУ ИЗДАНИЮ

Что побудило нас подготовить второе издание учебника? Во-первых, весь пятидесятитысячный тираж первого издания разошелся за неполные два года несмотря на внушительный объем книги и относительно специфическую область знаний, к которой она относится. Во-вторых, — это наш опыт общения со студентами и коллегами-преподавателями. В том, что учебник пользуется спросом у студентов, убеждает не только статистика продаж, но и специальные анкетные обследования, и информация, размещаемая на студенческих интернетовских сайтах (см. www.sachok.ru). Что касается вузовских преподавателей эконометрики, то прекрасную возможность общения с ними предоставила одному из авторов серия специально организованных для них семинаров в разных регионах России и бывших республик СССР, на которых отечественные и зарубежные специалисты (в их числе — С.А. Айвазян) представляли свои циклы лекций в рамках общей программы повышения квалификации преподавателей по эконометрике. С 1997-го по 2001 год такие семинары прошли в Москве (дважды), Санкт-Петербурге (дважды), Екатеринбурге, Нижнем Новгороде, Сочи, Владивостоке, Воронеже, Перми, Вильнюсе и других городах.

Второе издание выходит стереотипным, без серьезных доделок или изменений. Устранены лишь обнаруженные (к сожалению, в весьма большом количестве) опечатки и явные погрешности. Кроме того, мы предложили издательству выпустить учебник в двух томах: том 1 — «Теория вероятностей и прикладная статистика» (введение и главы 1–13 первого издания); том 2 — «Основы эконометрики» (главы 14–17 первого издания). Поскольку материал второго тома подготовлен полностью С.А. Айвазяном, этот том выходит только под его авторством. Практика использования студентами и преподавателями первого издания учебника подсказала нам, что двухтомный вариант книги может оказаться более технологичным и удобным в использовании.

Отметим, что со временем несколько трансформируются представления специалистов о самом предмете эконометрики, пополняется багаж его методов, смещаются акценты. Не со всеми такими представлениями, принятыми, скажем, в научных кругах США, согласны авторы этого учебника.

ка. Там, например, принято включать в продвинутые курсы (и учебники) по эконометрике «Теорию больших выборок» (или «Асимптотическую теорию»), «Непараметрические и полупараметрические методы принятия статистических решений», развернутое изложение метода максимального правдоподобия. С нашей точки зрения, вся эта тематика традиционно представлена в качестве разделов в других самостоятельных научных дисциплинах — теории вероятностей и математической статистике. В то же время важнейшие для эконометрического анализа *прикладные методы многомерной статистики* (дискриминантный и кластер анализы, главные компоненты и др.) по непонятным для нас причинам отсутствуют в эконометрических курсах и учебниках Северной Америки и Западной Европы.

Следует, однако, признать, что во втором томе предлагаемого издания представлены, конечно, далеко не все важнейшие разделы современной эконометрики (поэтому этот том и называется «О с н о в ы эконометрики»). В нем нет, например, *обобщенного метода моментов*, методов анализа *панельных данных*, раздела, посвященного *моделям с урезанными и цензурированными выборками*, недостаточно внимания уделено проблемам исследования стационарности временного ряда и, в связи с этим, *приемам коинтеграции и анализу единичных корней характеристического уравнения временного ряда*. Подобным образом расширенный вариант нашего второго тома составит, по существу, новый учебник, создание которого входит в ближайшие планы одного из авторов.

Наконец, о з а д а ч а х и у п р а ж е н е н и я х, которыми необходимо оснастить аудиторные занятия со студентами по прикладной статистике и эконометрике. Одновременно с выходом данного двухтомного издания издательство «Юнити-Дана» публикует нашу книгу «Прикладная статистика в задачах и упражнениях», которая является естественным дополнением (задачником) к первому тому. Аналогичное дополнение ко второму тому (задачник по эконометрике) один из авторов планирует представить в течение ближайшего года.

В заключение хотим поблагодарить профессора Гарвардского университета Дэйла Джоргенсона (Dale Jorgenson) за внимание к книге и очень полезные обсуждения, которые состоялись во время пребывания С.А. Айвазяна в Гарвардском университете в апреле 2001г. Мы благодарны также д-ру Джону Киммелу (John Kimmel) из Североамериканского отделения издательства «Шпрингер-Верлаг» за организацию рецензий западных специалистов на наш учебник. Мы искренне благодарны преподавателям статистики и эконометрики различных вузов России и Литвы — участникам упомянутых выше региональных семинаров по пре-

подаванию эконометрики за полезные обсуждения отдельных фрагментов учебника. Наконец, мы благодарны коллективам и руководству Центрального экономико-математического института Российской академии наук и Московского государственного университета экономики, статистики и информатики, плодотворная профессиональная среда существенно помогла нам в работе над учебником.

Москва, июнь 2001 г.

В в е д е н и е

The background of the title page features a series of concentric circles and a spiral pattern in a dark gray color, centered behind the main title text.

**Вероятностно-
Статистические
Методы в
Моделировании
Социально-
Экономических
Процессов и
Анализе данных**

ВВЕДЕНИЕ. ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКИЕ МЕТОДЫ В МОДЕЛИРОВАНИИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ПРОЦЕССОВ И АНАЛИЗЕ ДАННЫХ

В.1. Математико-статистический инструментарий экономических исследований

В.1.1. Назначение и составные части учебника

Если экономист строит рассуждения и выводы, *опираясь в своих модельных построениях на результаты конкретных измерений* интересующих его экономических, социально-экономических и (или) демографических показателей, то тем самым определяется **эконометрический подход к проблеме**, а инструментарий, которым он при этом пользуется, обязательно содержит те или иные **методы и модели прикладной математической статистики**. Ведь назначение эконометрики — придавать конкретное количественное выражение общим (качественным) закономерностям экономической теории на базе экономической статистики и с использованием средств математической статистики (подробнее определение эконометрики см. в п. 1.1 в томе 2).

Так вот, *предлагаемый вниманию читателя учебник как раз и посвящен описанию математико-статистического инструментария экономических исследований*. Почему же тогда не назвать этот учебник просто «Методы эконометрики»? По-существу можно было бы считать такое название этой книги вполне оправданным. Однако от такого решения авторов удержали два обстоятельства.

Первое обстоятельство связано с установившимся традиционным пониманием методов эконометрики, при котором они ограничивались

линейным регрессионным анализом (включающим классический и обобщенный методы наименьших квадратов, случаи гетероскедастичных и автокоррелированных регрессионных остатков, стохастических объясняющих переменных, а также использование фиктивных и инструментальных переменных), **методами и моделями анализа временных рядов** (так называемые AR-, MA-, ARMA- и ARIMA-модели, а в последние годы и модели коинтеграции, ARCH- и GARCH-модели) и, наконец, **системами одновременных уравнений**. Этот набор традиционных эконометрических методов описан во втором томе учебника.

Второе обстоятельство связано с первым и заключается в том, что по своему назначению, по своим инструментальным возможностям прикладная статистика — самостоятельная научная дисциплина, обслуживающая значительно более широкий класс реальных задач статистического анализа данных, чем тот, который традиционно непосредственно связывается с эконометрикой. К таким задачам в сфере социально-экономической теории и практики относятся, в частности:

- исследование динамики структуры состояний объектов (демографической или социальной структуры общества, структуры типологии потребительского поведения домашних хозяйств и т. п.);
- типологизация социально-экономических объектов (семей, фирм, предприятий, регионов, стран и т. п.);
- построение интегральных индикаторов качества или эффективности функционирования социально-экономической системы (уровня или качества жизни, качества населения, эффективности функционирования предприятия и т. п.);
- выявление скрытых (латентных) факторов, определяющих течение того или иного социально-экономического процесса;
- исследование и моделирование генезиса анализируемых статистических данных.

К традиционным экономическим задачам *применения* эконометрических методов принято относить (эта точка зрения представлена, например, в [Джонстон Дж., с. 16]) макро моделирование функционирования национальной экономики (агрегированное, не агрегированное, детализированное) или ее секторов. Традиционные приложения эконометрических методов на *микро*-уровне сводятся обычно к моделированию поведения потребителя, производителя и продавца, а также к моделированию некоторых процессов, происходящих на финансовых рынках.

Современные университетские учебные планы подготовки экономистов и экономистов-статистиков построены таким образом, чтобы обеспечивать непрерывность и преемственность обучения по **блокам дисци-**

плин:

- гуманитарному;
- экономическому;
- предметной статистики;
- математико-статистического инструментария;
- информационных технологий.

В качестве обязательных компонентов в **блоке математико-статистического инструментария** представлены курсы по *элементарным методам статистического анализа данных* (или дескриптивной статистике), *теории вероятностей*, *математической статистике*, *прикладной статистике* (или многомерным статистическим методам), *эконометрике*. Данный учебник посвящен взаимосвязному изложению всех упомянутых дисциплин блока математико-статистического инструментария.

В.1.2. Прикладная статистика

Нужно ли использовать этот термин или можно ограничиться более привычным понятием «математическая статистика»? Как соотносится прикладная статистика с другими статистическими дисциплинами, такими, как «математическая статистика», «анализ данных», «методы эконометрики»? Для обоснования правомерности и целесообразности рассмотрения прикладной статистики как самостоятельной научной дисциплины следует упомянуть, как минимум, о двух моментах.

Во-первых, до сих пор развитие теории, методологии и практики статистической обработки анализируемых данных шло по существу в двух параллельных направлениях. Одно из них представлено методами, *предусматривающими возможность вероятностной интерпретации* обрабатываемых данных и полученных в результате обработки статистических выводов. Именно эти методы и составляют содержание подавляющего большинства монографий и руководств по математической статистике. Другими словами, под методами математической статистики принято понимать лишь те методы статистической обработки исходных данных, разработка и использование которых апеллируют к вероятностной природе этих данных.¹ При этом развиваемый в рамках второго направления

¹ Такова ситуация, сложившаяся лишь *de facto*. Формально же, *de jure*, если исходить из определения Большого энциклопедического словаря (М., Большая Российская энциклопедия, 1997, с. 701), математическая статистика понимается более широко, а именно как «наука о математических методах систематизации и использования статистических данных для научных и практических выводов. Во многих своих разделах математическая статистика опирается на теорию вероятностей, позволяющую

весьма широкий и актуальный класс методов статистической переработки исходной информации, а именно вся совокупность тех методов, которые априори *не опираются на вероятностную природу обрабатываемых данных* (представителями методов такого типа являются, например, разнообразные методы *кластер-анализа, многомерного шкалирования, теории измерений* и др.), остается за общепринятыми рамками научной дисциплины «математическая статистика».

Во-вторых, специалисты, занимающиеся разработкой и конкретными применениями методов статистической обработки исходной информации, не могут игнорировать ту внушительную дистанцию, которая разделяет момент успешного завершения разработки собственно математического метода и момент получения результата от использования этого метода в решении конкретной практической задачи. В процессе прохождения этой трудной дистанции математику-прикладнику приходится, в частности:

- глубоко вникать в содержательную сущность задачи, адекватно «прилаживать» исходные модельные допущения (на которых строится любой математический метод) к выяснению сущности реальной задачи;
- решать задачу преобразования имеющейся исходной информации к стандартной (унифицированной) форме записи обрабатываемых статистических данных;
- разрабатывать *практически реализуемые* вычислительные алгоритмы и программное обеспечение с учетом специфики обрабатываемой статистической информации и возможностей имеющейся вычислительной техники.

Понятийный аппарат, методы и результаты, позволяющие проходить эту дистанцию, *вместе с этапом «прилаживания» и доработки необходимого математического инструментария и составляют* главное содержание прикладной статистики.

Таким образом, мы приходим к *определению прикладной статистики как самостоятельной научной дисциплины, разрабатывающей и систематизирующей понятия, приемы, математические методы и модели, предназначенные для организации сбора¹, стандартной записи, си-*

оценить надежность и точность выводов, делаемых на основании ограниченного статистического материала».

¹ Говоря об «организации сбора» статистических данных, мы имеем в виду лишь определение способа отбора подлежащих статистическому обследованию единиц (семей, предприятий, стран, пациентов и т. п.) из всей исследуемой совокупности (см. п. 6.1, а также в п. 9.1 описание типа 2). Мы не включаем сюда разработку *методологии измерителей анализируемых свойств* отображаемого объекта: эта работа предполагает *профессиональное* (экономическое, техническое, медицинское и т. п.) изуче-

стематизации и обработки статистических данных с целью их удобного представления, интерпретации и получения научных и практических выводов.

Для определения той же самой системы понятий, приемов, математических методов и моделей некоторые специалисты используют термин «анализ данных», понимаемый в расширительном толковании. Описанию методов прикладной статистики, выходящих за традиционные рамки методов эконометрики и одновременно наиболее актуальных для экономических и социально-экономических приложений, посвящен раздел III данного учебника.

В.1.3. Теория вероятностей и математическая статистика

Эти две дисциплины основные «поставщики» математического инструментария для прикладной статистики и эконометрики.

«Теория вероятностей — наука, позволяющая по вероятностям одних случайных событий находить вероятности других случайных событий, связанных каким-либо образом с первыми ... Можно также сказать, что теория вероятностей есть математическая наука, выясняющая закономерности, которые возникают при взаимодействии большого числа случайных факторов» (Математическая энциклопедия, М., Советская энциклопедия, 1976, т. 1, с. 655-656). Подчеркнем, что упомянутые в определении закономерности формулируются в терминах модельных соотношений.

Идеальной средой применимости теоретико-вероятностного способа рассуждения (и соответствующего математического аппарата) является ситуация, когда мы находимся в условиях *стационарного* (т. е. не изменяющегося во времени) действия некоторого *реального комплекса условий*, включающего в себя неизбежность «мешающего» влияния большого числа случайных (не поддающихся строгому учету и контролю) факторов, которые в свою очередь не позволяют делать полностью достоверные выводы о том, произойдет или не произойдет интересующее нас событие. При этом предполагается, что мы имеем принципиальную возможность (хотя бы мысленно реально осуществимую) многократного повторения нашего эксперимента или наблюдения в рамках того же самого реального комплекса условий. Именно такую ситуацию принято называть *условиями действия статистического ансамбля* или условиями соблюдения ста-

ние сущности задач, для решения которых требуется статистическая информация, а потому она относится к компетенции *предметной статистики* соответствующей области.

тистической однородности исследуемой совокупности (подробнее об этом см. п. В.2).

Необходимый для нормального понимания методов прикладной статистики и эконометрики минимум сведений из теории вероятностей приводится в разделе I данного учебника.

Одно из возможных определений *математической статистики* приведено в п. В.1.2 (см. сноску на с.26). Добавим к этому, что математическая статистика, являясь по отношению к прикладной статистике и эконометрике разработчиком и поставщиком существенной части используемого в них математического аппарата, полностью отстранена от таких функций этих прикладных дисциплин, как:

- «прилаживание» и доработка необходимого математического инструментария в соответствии со спецификой анализируемого класса реальных социально-экономических задач;
- разработка невероятностных (логико-алгебраических, оптимизационных и др.) методов анализа и моделирования реальных исходных данных, т.е. методов, не опирающихся на модельные допущения о вероятностной природе этих данных;
- преобразование разнообразных форм исходных социально-экономических данных с целью их удобного представления для дальнейшего анализа и моделирования;
- разработка практически реализуемых вычислительных алгоритмов и удобного программного обеспечения для используемых в ходе эконометрического анализа методов и моделей.

Раздел II учебника посвящен изложению основных результатов математической статистики, включая элементарные методы статистического анализа данных (методы *дескриптивной статистики*).

В.2. Теоретико-вероятностный способ рассуждения в прикладной статистике и эконометрике

В.2.1. Границы применимости теоретико-вероятностного способа рассуждения

В п. В.1.2 упоминалось о двух подходах к статистическому анализу данных: математико-статистическом, основанном на вероятностном способе рассуждения, и логико-алгебраическом. Ко второму подходу исследователь *вынужден* обращаться лишь тогда, когда условия сбора (регистра-

ции) исходных данных не укладываются в рамки так называемого **статистического ансамбля**, т.е. в *ситуацию, когда не имеется практической или хотя бы принципиально мысленно представимой возможности многократного тождественного воспроизведения основного комплекса условий, при которых производились измерения анализируемых данных*.

В условиях же статистического ансамбля исследователь имеет возможность воспользоваться классическими математико-статистическими методами обработки данных, когда для обоснования наилучшего выбора методов статистической переработки, итогового представления и интерпретации анализируемых данных он использует те или иные априорные сведения об их *случайной (стохастической)* природе. При этом мы исходим из того, что даже постулируемая нами тождественность воспроизведения основного комплекса условий эксперимента или наблюдения в большинстве реальных ситуаций (с учетом их сложности, множественности и частичной неизученности формирующих их факторов) *не избавляет нас от неконтролируемого (случайного) разброса* в самих результатах наблюдения. Так, даже практически идеально отлаженный станок автоматической линии не в состоянии производить абсолютно идентичные между собой (и заданному номиналу) изделия. В аналогичных условиях статистического ансамбля и соответствующего неконтролируемого разброса изучаемых показателей или событий мы окажемся, например, при изучении числа дефектных изделий в партиях заданного объема, отбираемых от массовой продукции и производимых в стационарном режиме производства, или при регистрации среднедушевого дохода семей, случайно отбираемых из некоторой однородной (в социальном, географическом и экономическом смысле) совокупности, и т. д.

Именно разнообразные математические модели таких скрытых закономерностей вместе с теоретическим и эмпирическим анализом их свойств и взаимоотношений и предоставляют исследователю *теория вероятностей и математическая статистика*.

Наиболее простые и убедительные примеры реальных ситуаций, подчиняющихся требованию статистической устойчивости (или укладывающихся в рамки статистического ансамбля), предоставляет нам область азартных игр.¹ Действительно, подбрасывая монету, бросая игральную

¹ Исторически именно эта область существенно стимулировала зарождение и развитие элементов теоретико-вероятностной научной дисциплины. Первые достаточно интересные результаты теории вероятностей принято связывать с работами Л. Пачоли («Сумма арифметики, геометрии, учения о пропорциях и отношениях», 1494 г.), Д. Кардано («Книга об игре в кости», 1526 г.) и Н. Тартальи («Общий трактат о числе и мере» 1556–1560 гг.). См. [Майстров Л. Е., с. 6, 25–37].

кость или вытягивая наугад карту из колоды и интересуясь при этом вероятностью осуществления события, заключающегося соответственно в появлении «герба», «шестерки» или «дамы пик», мы имеем все основания полагать, что

а) можно многократно повторить тот же самый эксперимент в тех же самых условиях;

б) наличие большого числа случайных факторов, характеризующих условия проведения каждого такого эксперимента, не позволяет делать полностью определенного (детерминированного) заключения о том, произойдет в результате данного эксперимента интересующее нас событие или не произойдет;

в) чем большее число однотипных экспериментов мы произведем, тем ближе будут подсчитанные по результатам экспериментов относительные частоты появления интересующих нас событий к некоторым постоянным величинам, называемым *вероятностными* для этих событий, а именно: относительная частота появления «герба» будет приближаться к $1/2$, выпадения «шестерки» — к $1/6$, а извлечения «дамы пик» (из колоды, содержащей 52 карты) — к $1/52$.

Очевидно, требования статистического ансамбля применительно к указанным выше трем типам экспериментов означают необходимость использования одной и той же (или совершенно идентичных) симметричной монеты, симметричной кости, а в последнем случае — необходимость возвращения извлеченной в предыдущем эксперименте карты в колоду и тщательного случайного перемешивания последней.

Соблюдение условий статистического ансамбля в более серьезных и сложных сферах человеческой деятельности — в экономике, в социальных процессах, в технике и промышленности, в медицине, в различных отраслях науки — это вопрос, требующий специального рассмотрения в каждом конкретном случае.

Оценивая специфику задач в различных областях человеческого знания с позиции соблюдения в них свойств а)–в) статистической устойчивости и принимая во внимание накопившийся опыт вероятностно-статистических приложений, можно условно разбить все возможные приложения на три категории.

К первой категории возможных областей применения — **категории высокой работоспособности вероятностно-статистических методов** — отнесем те ситуации, в которых свойства а)–в) статистической устойчивости исследуемой совокупности бесспорно имеют место либо нарушаются столь незначительно, что это практически не влияет на точность статистических выводов, полученных с использованием теоретико-

вероятностных моделей. Сюда (помимо упоминавшейся «игровой» области) могут быть отнесены отдельные разделы *экономики и социологии* и в первую очередь задачи, связанные с исследованием поведения объекта (индивидуума, семьи или другой социально-экономической или производственной единицы) как представителя большой однородной совокупности подобных же объектов. Традиционной областью эффективного использования вероятностно-статистического аппарата давно стала *демография*. Теоретико-вероятностные понятия являются основным языком в таких инженерных областях, как *теория надежности систем*, состоящих из очень большого числа элементов, и *теория выборочного контроля качества* продукции. В *медицине* вероятностно-статистический подход позволил ввести понятие факторов риска развития основных хронических заболеваний и провести количественное изучение их влияния, способствуя тем самым большей индивидуализации, а значит, и эффективности профилактики и лечения. Результаты специальных вероятностно-статистических исследований выявили, что вероятность дожить до определенного возраста подвержена не слишком значительным колебаниям (в зависимости от условий жизни). Эти результаты и послужили основой составления так называемых *таблиц выживаемости*, в определенной мере и в определенном смысле (а именно, в среднестатистическом, но не в индивидуальном, конечно!) поколебавшим известное древнее изречение «никто не знает часа своей смерти». Вероятностно-статистический способ рассуждения играет видную роль в исследованиях, проводимых в современной физике (в первую очередь в *статистической физике*) и в классической механике (в *статистической теории газов*).

Отметим одну важную общую черту, характеризующую подавляющее большинство задач перечисленных выше областей человеческой деятельности, в которых оказывается правомерным и эффективным применение вероятностно-статистических методов. Речь идет о *существенной многомерности* обрабатываемой информации, характеризующей исследуемые явления или объекты, т.е. о ситуациях, когда состояние или поведение каждого из этих объектов в любой фиксированный момент времени описывается *набором* соответствующих показателей. Среди этих показателей могут быть как количественные (среднедушевой доход в семье, размер семьи, объем валовой продукции предприятия и т.д.), так и не количественные, т.е. ранговые (классификация специалиста, сравнительная характеристика жилищных условий) и классификационные, или номинальные (профессия, национальность; пол, причины миграции и т.п.). Все эти показатели находятся в сложной взаимосвязи друг с другом. Именно в таких ситуациях принято говорить о многомерности исследуемой схемы, а

исследователю приходится обращаться к методам *многомерного статистического анализа*.

Ко второй категории возможных областей применения — категории **допустимых вероятностно-статистических приложений** — отнесем ситуации, характеризующиеся весьма значительными нарушениями требования *сбътрания неизменными условий эксперимента* (вторая половина требования а)) и вытекающими отсюда отклонениями от требования в). Характерной формой такого рода отклонений от условий статистического ансамбля является объединение в одном ряду наблюдений (подлежащих обработке) различных порций исходных данных, зарегистрированных в разных условиях (в разное время или в разных совокупностях). К этой же категории приложений можно отнести определенный класс задач, связанных с анализом *коротких временных рядов*, зарегистрированных в условиях, практически исключающих возможность статистической фиксации сразу нескольких эмпирических реализаций исследуемого временного ряда на одном и том же временном интервале. Использование вероятностно-статистических методов обработки в этом случае допустимо, но должно сопровождаться пояснениями о несовершенстве и приближенном характере получаемых при этом выводов (например, не следует слишком доверять в подобных ситуациях различным числовым характеристикам степени достоверности этих выводов, т.е. *доверительным вероятностям, уровням значимости* критерия и т.п.) и по возможности должно дополняться другими методами научного анализа.

И наконец, к *третьей категории* задач статистической обработки исходных данных — категории **недопустимых вероятностно-статистических приложений** — следует отнести ситуации, характеризующиеся либо принципиальным неприятием главной идеи понятия статистического ансамбля — *массовости* исследуемой совокупности (т.е. конкретной бессодержательностью идеи многократного повторения одного и того же эксперимента в неизменных условиях, сформулированной в требовании а)), либо полной *детерминированностью* изучаемого явления, т.е. — отсутствием «мешающего» влияния множества случайных факторов (нарушение требования б)). В подобных ситуациях исследователь должен пользоваться методами анализа данных (см. [Диде Э. и др.]) и не должен претендовать на вероятностную интерпретацию обрабатываемых данных и получаемых в результате их обработки выводов.

Строгих математических методов, позволяющих точно определять, находимся ли мы в условиях статистического ансамбля, не существует: любая вероятностная модель, так же как и любая математическая модель вообще, есть лишь некоторая аппроксимация исследуемой реальной дей-

ствительности. Можно говорить лишь о ситуациях, очевидно укладывающихся в рамки статистического ансамбля (бросание монеты, игральной кости; контроль продукции массового производства, работающего в отлаженном стационарном режиме, и т. п.), укладывающихся в эти рамки лишь приблизительно, с оговорками, и явно не соответствующих условиям статистического ансамбля. Однако даже с последней категорией ситуаций (названных у нас категорией недопустимых вероятностно-статистических приложений) нет полной ясности. Так, например, с позиций статистического ансамбля события типа «в 2050 г. начнется война между странами А и В» явно не относятся к сфере возможных применений вероятностно-статистических методов — налицо нарушение требования а)!

Однако существует концепция так называемых *субъективных вероятностей*, в рамках которой оказывается правомерным говорить и о вероятности таких событий. Для этого следует прибегнуть к помощи экспертов и вместо действительной многократной реализации интересующего нас эксперимента в одних и тех же условиях ограничиться воображаемой прогонкой исследуемой ситуации «через сознание» многих экспертов. При этом, очевидно, *эксперт интерпретируется как некий измерительный прибор*, работающий со случайной ошибкой. Точность работы этого «измерительного прибора», т. е. точность «прочтения» (в сознании эксперта) исхода интересующего нас события в будущем, очевидно, зависит как от степени объективного влияния «мешающих» случайных/факторов (т. е. от степени временной отдаленности интересующего нас момента времени, общей сложности ситуации и т. п.), так и от степени осведомленности, компетентности и других субъективных качеств самого эксперта. Мы не собираемся здесь вмешиваться в спор между субъективистской и классической вероятностными концепциями. Останемся на той точке зрения, что единственным объективным судьей в подобных вопросах может быть лишь критерий практики. В этой связи уместно напомнить читателю следующие слова Ф. Энгельса из «Анти-Дюринга»: «... Математика, вообще столь строго нравственная, совершила грехопадение: она вкусила от яблока познания, и это открыло ей путь к гигантским успехам, но вместе с тем и к заблуждениям. Девственное состояние абсолютной значимости, неопровержимой доказанности всего математического навсегда ушло в прошлое; наступила эра разногласий, и мы дошли до того, что большинство людей дифференцирует и интегрирует не потому, что люди понимают, что они делают, а просто потому, что верят в это, так как до сих пор результат всегда получался правильным»¹

¹ Маркс К., Энгельс Ф. Соч., т. 20, с. 89.

Резюмируя содержащееся в данном пункте обсуждение сущности, назначения и границ применимости теории вероятностей, мы можем через полтора века после замечательного французского ученого Лапласа повторить, более обоснованно и убедительно, что «замечательно, что наука, которая начала с изучения игр, возвысилась до наиболее важных предметов человеческого знания».²

В.2.2. Что дает объединение теоретико-вероятностного и статистического способов рассуждения?

Теперь попытаемся на простом примере показать преимущества вероятностно-статистического (или математико-статистического) способа принятия решения, его отличие как от чисто статистического, так и от чисто теоретико-вероятностного.

Статистический способ принятия решения. Пусть читатель представит себя наблюдающим за игрой двух лиц в кости, происходящей по следующим правилам. Производится 4 последовательных бросания игральной кости. Игрок *A* получает одну денежную единицу от игрока *B*, если в результате этих четырех бросаний хотя бы один раз выпало шесть очков (назовем этот исход «шесть»), и платит одну денежную единицу игроку *B* в противном случае (назовем этот исход «не шесть»). После ста туров читатель должен сменить одного из игроков, причем он имеет право выбрать ситуацию, на которую он будет ставить свою денежную единицу в следующей серии туров: за появление хотя бы одной «шестерки» или против. Как правильно осуществить этот выбор?

Статистический способ решения этой задачи диктуется обычным здравым смыслом и заключается в следующем. Пронаблюдав сто туров игры предыдущих партнеров и подсчитав относительные частоты их выигрыша, казалось бы, естественно поставить на ту ситуацию, которая чаще возникала в процессе игры. Например, было зафиксировано, что в 52 партиях из 100 выиграл игрок *B*, т. е. в 52 турах из 100 «шестерка» не выпадала ни разу при четырехкратном выбрасывании кости (соответственно в остальных 48 партиях из ста осуществлялся исход «шесть»). Следовательно, делает вывод читатель, применивший статистический способ рассуждения, выгоднее ставить на исход «не шесть», т. е. на тот исход, относительная частота появления которого ($\hat{p}$) равна 0,52 (больше половины).

² См.: Oeuvres complètes de Laplace. T. 7: «Théorie analytique des probabilités». Paris, Gauthier-Villars, 1886, p/ CLII.

Теоретико-вероятностный способ решения. Этот способ основан на определенной *математической модели* изучаемого явления: полагая кость правильной (т.е. симметричной), а следовательно, принимая шансы выпадения любой грани кости при одном бросании равными между собой (другими словами, относительная частота, или вероятность, выпадения «единицы» равна относительной частоте выпадения «двойки» и т.д... равна относительной частоте выпадения «шестерки» и равна $1/6$), можно подсчитать вероятность P {«не шесть»} осуществления ситуации «не шесть», т.е. вероятность события, заключающегося в том, что при четырех последовательных бросаниях игральной кости ни разу не появится «шестерка». Этот расчет основан на следующих фактах, вытекающих из принятой нами математической модели. Вероятность *не выбросить* шестерку при *одном бросании* кости складывается из шансов появиться в результате одного бросания «единице», «двойке», «тройке», «четверке» и «пятерке» и, следовательно, составляет (в соответствии с определением вероятности любого события, см. п. 1.1) $5/6$. Затем используем теорему умножения вероятностей (см. п. 1.1.3), в соответствии с которой вероятность наступления нескольких независимых событий равна произведению вероятностей этих событий. В нашем случае мы рассматриваем факт наступления четырех независимых событий, каждое из которых заключается в невыпадении «шестерки» при одном бросании и имеет вероятность осуществления, равную $5/6$. Поэтому

$$p = P \{ \ll \text{не шесть} \gg \} = \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \\ = \frac{625}{1296} = 0,482.$$

Как видно, вероятность ситуации «не шесть» оказалась меньше половины, следовательно, шансы ситуации «шесть» предпочтительнее (соответствующая вероятность равна: $1 - 0,482 = 0,518$). А значит, читатель, использовавший теоретико-вероятностный способ рассуждения, придет к диаметрально противоположному по сравнению с читателем со статистическим образом мышления решению и будет ставить в игре на ситуацию «шесть».

Вероятностно-статистический (или математико-статистический) способ принятия решения. Этот способ как бы синтезирует инструментарий двух предыдущих, так как при выработке с его помощью окончательного вывода используются и накопленные в результате наблюдения за игрой *исходные статистические данные* (в виде относительных частот появления ситуаций «шесть» и «не шесть», которые,

как мы помним, были равны соответственно 0,48 и 0,52), и *теоретико-вероятностные модельные соображения*. Однако модель, принимаемая в данном случае, менее жестка, менее ограничена, она как бы *настраивается на реальную действительность, используя для этого накопленную статистическую информацию*. В частности, эта модель уже не постулирует правильность используемых костей, допуская, что центр тяжести игральной кости может быть и смещен некоторым специальным образом. Характер этого смещения (если оно есть) должен как-то проявиться в тех исходных статистических данных, которыми мы располагаем. Однако читатель, владеющий вероятностно-статистическим образом мышления, должен отдавать себе отчет в том, что полученные из этих данных величины относительных частот исходов «шесть» и «не шесть» дают лишь некоторые *приближенные оценки* истинных (теоретических) шансов той и другой ситуации: ведь подбрасывая, скажем, 10 раз даже идеально симметричную монету, мы можем случайно получить семь выпадений «гербов»; соответственно относительная частота выпадения «герба», подсчитанная по этим результатам испытаний, будет равна 0,7; но это еще не значит, что истинные (теоретические) шансы (вероятности) появления «герба» и другой стороны монеты оцениваются величинами соответственно 0,7 и 0,3 — эти вероятности, как мы знаем, равны 0,5. Точно так же наблюдаемая нами в серии из ста игровых туров относительная частота исхода «не шесть» (равная 0,52) может отличаться от истинной (теоретической) вероятности того же события и, значит, может не быть достаточным основанием для выбора этой ситуации в игре! Весь вопрос в том, насколько сильно может отличаться наблюдаемая (в результате осуществления n испытаний) относительная частота $\hat{p}_n$ интересующего нас события от истинной вероятности p появления этого события и как это отличие, т.е. погрешность $\hat{p}_n - p$, зависит от числа n имеющихся в нашем распоряжении наблюдений? (интуитивно ясно, что чем дольше мы наблюдали за игрой, т.е. чем больше общее число n использованных нами наблюдений, тем больше доверия заслуживают вычисленные нами эмпирические относительные частоты $\hat{p}_n$, т.е. тем меньше их отличие от неизвестных нам истинных значений вероятностей p .) Ответ на этот вопрос можно получить в нашем случае, если воспользоваться рядом *модельных соображений*: а) интерпретировать реализацию любого числа игровых партий как последовательность так называемых испытаний Бернулли, что означает, что результат каждого тура никак не зависит от результатов предыдущих туров, а неизвестная нам вероятность p осуществления ситуации «не шесть» остается одной и той же на протяжении всех туров игры; б) использовать тот факт, что поведение случайно меняющейся (при повторе-

ниях эксперимента) погрешности $\Delta_n = \hat{p}_n - p$ приближенно описывается *законом нормального* распределения вероятностей со средним значением, равным нулю, и дисперсией, равной $p(1-p)/n$ (см. п. 3.1.5).

Эти соображения, в частности, позволяют оценить абсолютную величину погрешности $|\Delta_n|$, которую мы можем допустить, заменяя неизвестную величину p вероятности интересующего нас события (в нашем случае — исхода «не шесть») относительной частотой $\hat{p}_n$ этого события, зафиксированной в серии из n испытаний (в нашем случае $n = 100$, а $\hat{p}_{100} = 0,25$). Если же мы смогли численно оценить абсолютную величину возможной погрешности Δ_n , то естественно применить следующее правило принятия решения: если относительная частота $\hat{p}_n$ появления исхода «не шесть» больше половины и продолжает превосходить 0,5 после вычитания из нее возможной погрешности $|\Delta_n|$, то выгоднее ставить на «не шесть»; если относительная частота $\hat{p}_n$ меньше половины и продолжает быть меньше 0,5 после прибавления к ней возможной погрешности $|\Delta_n|$, то выгоднее ставить на «шесть»; в других случаях у наблюдателя нет оснований для статистического вывода о преимуществах того или иного выбора ставки в игре (т.е. надо либо продолжить наблюдения, либо участвовать в игре с произвольным выбором ставки, ожидая, что это не может привести к сколь-нибудь ощутимому выигрышу или проигрышу).

Приближенный подсчет максимально возможной величины этой погрешности, опирающийся на модельное соображение б) (т.е. теорему Муавра-Лапласа, см. п. 4.3), дает в рассматриваемом примере, что с практической достоверностью, а именно с вероятностью 0,95, справедливо неравенство

$$|\hat{p}_n - p| \leq 2\sqrt{\frac{p(1-p)}{n}}. \quad (\text{В.1})$$

Возведение (В.1) в квадрат и решение получившегося квадратного неравенства относительно неизвестного параметра p дает

$$\begin{aligned} \frac{\hat{p}_n + \frac{2}{n}}{1 + \frac{4}{n}} - \frac{2\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{1}{n^2}}}{1 + \frac{4}{n}} &\leq p \\ &\leq \frac{\hat{p}_n + \frac{2}{n}}{1 + \frac{4}{n}} + \frac{2\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{1}{n^2}}}{1 + \frac{4}{n}}, \end{aligned}$$

или, с точностью до величин порядка малости выше, чем $1/\sqrt{n}$:

$$\hat{p}_n - 2\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \leq p \leq \hat{p}_n + 2\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}. \quad (\text{В.2})$$

В данном случае (при $\hat{p}_n = 0,52$ и $n = 100$) получаем

$$|\Delta_n| \approx 2\sqrt{\hat{p}_n(1 - \hat{p}_n)}/\sqrt{n} = 2\sqrt{0,52 \cdot (1 - 0,52)}/\sqrt{100} \approx 0,10.$$

Следовательно,

$$0,52 - 0,10 \leq p \leq 0,52 + 0,10.$$

Таким образом, имеющиеся в нашем распоряжении наблюдения за исходами ста партий дают нам основания лишь заключить, что интересующая нас неизвестная величина вероятности исхода «не шесть» на самом деле может быть любым числом из отрезка $[0,42; 0,62]$, т. е. может быть как величиной, меньшей 0,5 (и тогда надо ставить в игре на ситуацию «шесть»), так и величиной, большей 0,5 (и тогда надо ставить в игре на ситуацию «не шесть»).

Иначе говоря, читатель, воспользовавшийся вероятностно-статистическим способом решения задачи, вынужден будет прийти в данном случае к более осторожному выводу: ста партий в качестве исходного статистического материала оказалось недостаточно для вынесения надежного заключения о том, какой из исходов игры является более вероятным. Отсюда решение: либо продолжить роль «зрителя» до тех пор, пока область возможных значений для вероятности p , полученная из оценок вида (В.2), не окажется целиком лежащей левее или правее 0,5, либо вступить в игру, оценивая ее как близкую к «безобидной», т. е. к такой, в которой в длинной серии туров практически останешься «при своих».

Приведенный пример иллюстрирует роль и назначение теоретико-вероятностных и математико-статистических методов, их взаимоотношения. Если *теория вероятностей предоставляет исследователю набор математических моделей*, предназначенных для описания закономерностей в поведении реальных явлений или систем, функционирование которых происходит под влиянием большого числа взаимодействующих случайных факторов, то *средства математической статистики позволяют подбирать среди множества возможных теоретико-вероятностных моделей ту, которая в определенном смысле наилучшим образом соответствует имеющимся в распоряжении исследователя статистическим данным*, характеризующим реальное поведение конкретной исследуемой системы.

В.3. Вероятностно-статистическая (эконометрическая) модель как частный случай математической модели

В.3.1. Математическая модель

Математическая модель — это абстракция реального мира, в которой интересующие исследователя отношения между реальными элементами заменены подходящими отношениями между математическими категориями. Эти отношения, как правило, представлены в форме уравнений и (или) неравенств между показателями (переменными), характеризующими функционирование моделируемой реальной системы. Искусство построения математической модели состоит в том, чтобы совместить как можно большую лаконичность в ее математическом описании с достаточной точностью модельного воспроизводства именно тех сторон анализируемой реальности, которые интересуют исследователя.

Выше, анализируя в п. В.2.2 взаимоотношения чисто статистического, чисто теоретико-вероятностного и смешанного — вероятностно-статистического способа рассуждения, мы уже пользовались простейшими моделями, а именно:

- *статистической частотной моделью* интересующего нас случайного события, заключающегося в том, что в результате четырех последовательных бросаний игральной кости ни разу не выпадет «шестерка»; оценив по предыстории *относительную частоту* $\hat{p} = 0,52$ этого события и приняв ее за *вероятность появления этого события в будущем ряду испытаний*, мы, тем самым, используем модель случайного эксперимента с известной вероятностью его исхода;

- *теоретико-вероятностной моделью последовательности испытаний Бернулли* (см. п. 3.1.1), которая никак не связана с использованием результатов наблюдений (т. е. со статистикой); для подсчета вероятности интересующего нас события достаточно принятия гипотетического допущения о том, что используемая игральная кость идеально симметрична; тогда в соответствии с моделью серии независимых испытаний и справедливой, в рамках этой модели, теоремой умножения вероятностей подсчитывается интересующая нас вероятность по формуле (В.1);

- *вероятностно-статистической моделью*, интерпретирующей оцененную в чисто статистическом подходе относительную частоту $\hat{p}$ как некую случайную величину (см. п. 2.1), поведение которой подчиняется правилам, определяемым так называемой теоремой Муавра-Лапласа; при построении этой модели были использованы как теоретико-вероятностные

понятия и правила, так и статистические приемы, основанные на результатах наблюдений.

Обобщая этот пример, можно сказать, что:

- **вероятностная модель** — это математическая модель, имитирующая механизм функционирования *гипотетического* (не конкретного) реального явления (или системы) стохастической природы; в нашем примере гипотетичность относилась к свойствам игровой кости: она должна была быть идеально симметричной;
- **вероятностно-статистическая модель** — это вероятностная модель, значения отдельных характеристик (параметров) которой оцениваются по результатам наблюдений (исходным статистическим данным), характеризующим функционирование моделируемого *конкретного* (а не гипотетического) явления (или системы).

Вероятностно-статистическая модель, описывающая механизм функционирования экономической или социально-экономической системы, называется **эконометрической**. Если же речь идет о любой математической модели, описывающей механизм функционирования некоей гипотетической экономической или социально-экономической системы, то такую модель принято называть **экономико-математической** или просто **экономической**.

В качестве примера *экономической модели* рассмотрим простейший (идеализированный) вариант так называемой «**паутинной модели**», которая описывает процесс формирования спроса и предложения определенного товара или вида услуг на конкурентном рынке. Речь идет о формализации экономического закона спроса и предложения, гласящего: *количество товара, которое можно продать на рынке (т.е. спрос) изменяется в направлении, противоположном изменению его цены; количество товара, которое продавцы доставляют на рынок (т.е. предложение) изменяется в том же направлении, что и цена; при этом реальная рыночная цена складывается на уровне, при котором спрос и предложение равны друг другу (т.е. находятся в равновесии)*.¹

Займемся математической формализацией этих положений. Пусть x_t (ден. ед.) — цена товара в «момент времени» t . И пусть $y_t^{(n)}$ и $y_t^{(c)}$ — количество товара, соответственно предложенного и купленного («спрошенного») на рынке в тот же момент времени t . Тогда, с учетом одного такта времени, необходимого производителям-продавцам на то, чтобы «среагировать» на цену x , можно математически сформулировать приведенные

¹ Первым, кто попытался математически сформулировать этот закон, был А. Курно (Cournot A., *Recherches sur les principes mathématiques de la théorie des richesses*, Paris, 1838).

выше общие закономерности в виде:

$$\begin{cases} y_t^{(n)} = f(x_{t-1}); \\ y_t^{(c)} = g(x_t); \\ \lim_{t \rightarrow \infty} f(x_{t-1}) = \lim_{t \rightarrow \infty} g(x_t); \end{cases}$$

$$\lim_{t \rightarrow \infty} x_t = \bar{x}_q,$$

где $f(x)$ — некоторая монотонно возрастающая, а $g(x)$ — монотонно убывающая функции от аргумента x (т.е. от цены).

Математические соотношения, отражающие закон спроса-предложения, могут быть проиллюстрированы рисунком В.1

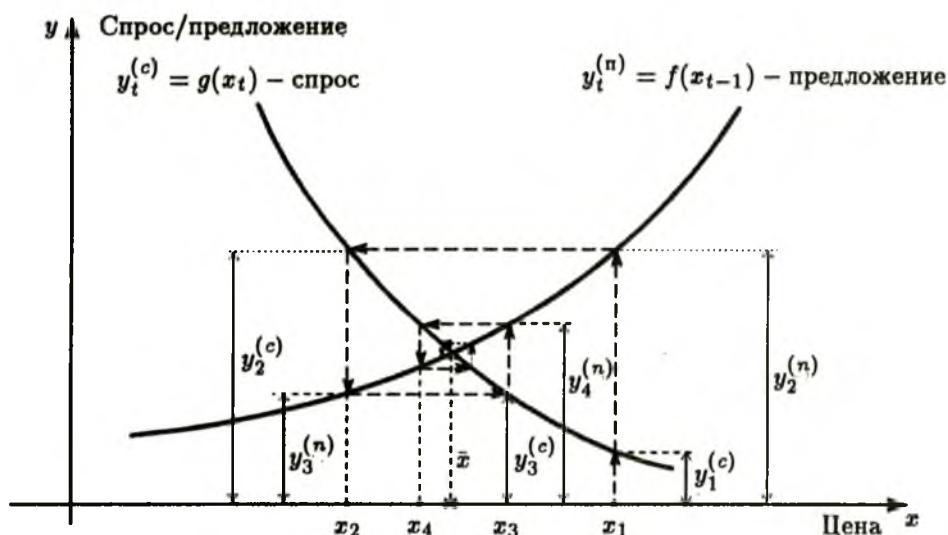


Рис. В.1. График процесса формирования спроса-предложения («паутинная» модель)

Из рисунка видно, что процесс формирования равновесной цены начался с назначения в 1-й (начальный) момент времени цены на уровне x_1 . Производитель-продавец отреагировал на это в следующий (2-й) момент времени величиной предложения, равной $y_2^{(n)} = f(x_1)$, в то время как спрос на этот товар сформировался всего на уровне $y_1^{(c)} = g(x_1)$. Заметное превышение предложения над спросом привело к понижению цены в следующий (2-й) момент времени до уровня x_2 . Это сразу отразилось на предложении в следующий (3-й) момент времени: оно снизилось до

$y_3^{(n)} = f(x_2)$. Зато спрос резко подскочил и составил во 2-й момент времени величину $y_2^{(c)} = g(x_2)$, и т. д., и т. п. Продолжение этого процесса обозначено траекторией (она индексирована стрелками), которая сходится *паутинообразно* к *точке равновесия* — к точке пересечения кривых $g(x)$ и $f(x)$.

Реалистическая модель закона спроса-предложения, конечно, сложнее. В частности, $y^{(n)}$ и $y^{(c)}$ зависят не только от цены x , поскольку *связь между $y^{(n)}$ и $y^{(c)}$, с одной стороны, и ценой x — с другой, носит не детерминированный, а стохастический характер* (о подобных зависимостях см. гл. 10, 11, а также гл. 2 во втором томе учебника).

Наконец, для того, чтобы эта модель превратилась из *экономической в эконометрическую*, следует говорить *не вообще* о законе спроса-предложения, а о *конкретном его действии в данном месте, данное время и применительно к данному конкретному товару (или виду услуг)*. Соответственно *конкретизация вида функций $f(x)$ и $g(x)$ должна производиться на базе исходных статистических данных о значениях $x_t, y_t^{(n)}$ и $y_t^{(c)}$ за ряд тактов времени (т. е. для $t = 1, 2, \dots, n$)*.

В.3.2. Основные этапы вероятностно-статистического моделирования

Построение и экспериментальная проверка (верификация) вероятностно-статистической модели обычно основаны на одновременном использовании информации двух типов:

(а) *априорной информации* о природе и содержательной сущности анализируемого явления, представленной, как правило, в виде тех или иных теоретических закономерностей, ограничений, гипотез;

(б) *исходных статистических данных*, характеризующих процесс и результаты функционирования анализируемого явления или системы.

Можно выделить следующие основные этапы вероятностно-статистического моделирования.

1-й этап (постановочный) включает в себя определение: конечных прикладных целей моделирования; набора факторов и показателей (переменных), описание взаимосвязей между которыми нас интересует; наконец, роли этих факторов и показателей — какие из них, в рамках поставленной конкретной задачи, можно считать *входными* (т. е. полностью или частично регулируемыми или хотя бы легко поддающимися регистрации и прогнозу; подобные факторы несут смысловую нагрузку *объясняющих* в модели), а какие — *выходными* (эти факторы обычно трудно поддаются непосредственному прогнозу; их значения формируются как бы в процессе

функционирования моделируемой системы, а сами факторы несут смысловую нагрузку *объясняемых*).

2-й этап (априорный, предмодельный) состоит в предмодельном анализе содержательной сущности моделируемого явления, формировании и формализации имеющейся априорной информации об этом явлении в виде ряда гипотез и исходных допущений (последние должны быть подкреплены теоретическими рассуждениями о механизме изучаемого явления или, если возможно, экспериментальной проверкой).

3-й этап (информационно-статистический) посвящен сбору необходимой статистической информации, т. е. регистрации значений участвующих в описании модели факторов и показателей на различных временных и (или) пространственных тактах функционирования моделируемой системы.

4-й этап (спецификация модели) включает в себя непосредственный вывод (опирающийся на принятые на 2-м этапе гипотезы и исходные допущения) *общего вида* модельных соотношений, связывающих между собой интересующие нас входные и выходные переменные. Говоря об общем виде модельных соотношений, мы имеем в виду то обстоятельство, что на данном этапе будет определена *лишь структура модели*, ее символическая аналитическая запись, в которой наряду с известными числовыми значениями (представленными в основном исходными статистическими данными) будут присутствовать величины, содержательный смысл которых определен, а числовые значения — нет (их обычно называют параметрами модели, неизвестные значения которых подлежат статистическому оцениванию).

5-й этап (идентифицируемость и идентификация модели) предназначен для проведения статистического анализа модели с целью «настройки» значений ее неизвестных параметров на те исходные статистические данные, которыми мы располагаем. При реализации этого этапа «модельер» должен сначала ответить на вопрос, *возможно ли в принципе однозначно восстановить значения неизвестных параметров модели* по имеющимся исходным статистическим данным при принятой на 4-м этапе структуре (способе спецификации) модели. Это составляет так называемую *проблему идентифицируемости* модели. А затем, после положительного ответа на этот вопрос, необходимо решить уже *проблему идентификации* модели, т. е. предложить и реализовать математически корректную процедуру оценивания неизвестных значений параметров модели по имеющимся исходным статистическим данным. Если проблема идентифицируемости решается отрицательно, то возвращаются к 4-му этапу и вносят необходимые коррективы в решение задачи спецификации модели.

6-й этап (верификация модели) заключается в использовании различных процедур сопоставления модельных заключений, оценок, следствий и выводов с реально наблюдаемой действительностью. Этот этап называют также этапом статистического анализа точности и адекватности модели. При пессимистическом характере результатов этого этапа необходимо возвратиться к этапу 4, а иногда и к этапу 1.

Построение и анализ модели могут быть основаны только на априорной информации (а) и не предусматривать проведения этапов 3 и 5. *Тогда модель не является вероятностно-статистической (эконометрической) при моделировании экономических закономерностей*. Более подробно о сущности и специфике именно эконометрического моделирования речь будет идти в гл. 1–3 второго тома учебника.

В.3.3. Моделирование механизма явления вместо формальной статистической фотографии

Обращаем внимание читателя на ключевую роль успешного проведения 2-го этапа в общей оценке степени реалистичности и работоспособности построенной модели.

Другими словами, адекватность и соответственно эффективность модели будут решающим образом зависеть от того, насколько глубоко и профессионально был проведен анализ *реальной сущности* изучаемого явления при формировании априорной информации (т. е. в рамках второго этапа). Поясним этот тезис. При вероятностно-статистическом моделировании и, в частности, на этапе формирования априорной информации о физической природе реального механизма преобразования входных показателей в выходные (результатирующие) *какая-то часть этого механизма остается скрытой от исследователя* (именно об этой части принято в соответствии с обиходной кибернетической терминологией говорить как о «черном ящике»). Чем большее профессиональное знание механизма исследуемого явления продемонстрирует исследователь, тем меньше будет доля «черного ящика» в общей логической схеме моделирования и тем работоспособнее и точнее будет построенная модель. Вероятностно-статистическое моделирование, полностью основанное на логике «черного ящика», позволяет получить исследователю лишь как бы *мгновенную статистическую фотографию* анализируемого явления, в общем случае непригодную, например, для целей прогнозирования. Напротив, моделирование, опирающееся на глубокий профессиональный анализ природы изучаемого явления, позволяет в значительной мере теоретически обосновать общий вид конструируемой модели, что дает основание к ее широко-

му и правомерному использованию в прогнозных расчетах. Поясним это на примере из п. 8.6.1.

Пусть целью нашего исследования является лаконичное (*параметризованное с помощью модели*) описание функции плотности анализируемой случайной величины (заработной платы наугад выбранного из общей генеральной совокупности работника) по исходным данным, представленным случайной выборкой работников $x_1, x_2, \dots, x_{750}$ объема $n = 750$. Игнорируя экономические закономерности формирования искомого закона распределения, т. е. руководствуясь *формальным* подходом наилучшей мгновенной статистической фотографии, мы должны были бы запастись достаточно богатым классом модельных плотностей (например, классом кривых Пирсона [Кендалл М. Дж., Стьюарт А., 1966]) и, перебирая эти модели (с одновременной статистической оценкой участвующих в их записи параметров методами, описанными в гл. 7), найти такую функцию плотности, которая наилучшим в определенном смысле образом (например, в смысле критерия «хи-квадрат» Пирсона, см. п. 8.6.1) аппроксимирует поведение имеющейся у нас эмпирической плотности (см. изображение соответствующей гистограммы на рис. 8.2). На этом пути в результате расширения запаса гипотетичных модельных плотностей можно добиться очень высокой точности аппроксимации, вплоть до повторения модельной функцией неожиданных провалов гистограммы, подобных тем, которые мы имеем на 14-м и 15-м интервалах группирования на рис. 8.2. Однако, поступая таким образом, мы добиваемся лишь кажущегося хорошего результата, в чем легко можно убедиться, попробовав применить выявленный модельный закон к описанию эмпирической плотности, построенной по *другой выборке, извлеченной из той же самой совокупности*. В подавляющем большинстве случаев выявленная ранее модельная плотность оказывается непригодной для описания распределительных закономерностей, наблюдаемых в другой выборке. Следовательно, для этой выборки нужно строить *другую модель*, а значит, и само моделирование практически теряет смысл, так как главное назначение модели — распространение закономерностей, подмеченных в выборке, на всю генеральную совокупность (что и является основой решения задач планирования, прогноза, диагностики).

В качестве альтернативного рассмотрим подход, предусматривающий тщательный предмодельный профессиональный анализ локальных закономерностей, в соответствии с которыми формируется закон распределения заработной платы. Эти закономерности (мультипликативный характер редукации труда, принцип оплаты по труду, постоянство относительного варьирования заработной платы при переходе от работников одной категории сложности труда к другой и т. п., см. [Айвазян С. А., 1980]

позволяют уже на следующем, третьем этапе моделирования *теоретически* (т.е. без апелляции к имеющейся у нас эмпирической функции плотности) обосновать выбор класса моделей, в пределах которого мы должны оставаться при подборе искомой модельной плотности. В рассмотренном примере таким классом был класс логарифмически-нормальных распределений (см. п. 3.1.6). После этого мы переходим к статистическому оцениванию параметров, участвующих в записи законов этого класса, т.е. переходим к четвертому этапу.

Модель, полученная таким образом, как правило, несколько хуже (по формальным критериям), чем предыдущая, аппроксимирует эмпирическую плотность, построенную *по данной конкретной выборке*. Однако в отличие от модели, полученной в результате формальной статистической подгонки экспериментальных данных под одну из теоретических кривых, она останется устойчивой, *инвариантной по отношению к смене выборки*, т.е. она одинаково хорошо может описывать характер распределения, наблюдаемого в различных выборках из одной и той же генеральной совокупности. А если все-таки моделирование, идущее от более или менее бесспорных (быть может, частично подтвержденных экспериментом) исходных предпосылок о физической природе изучаемого явления, дает результаты, плохо согласующиеся с реальной действительностью? Причина этого (при условии аккуратного проведения третьего и четвертого этапов) одна: плохое соблюдение на практике всех (или части) принятых при моделировании в качестве априорных допущений исходных предпосылок. Оценка же этого явления может быть двоякой: если заложенные в основание модели исходные допущения признаются специалистами объективными закономерностями, в соответствии с которыми *должен функционировать* механизм исследуемого явления, то следует искать и устранять причины, приведшие к нарушению этих закономерностей, *в самой моделируемой реальности*; если же принятые допущения были результатом *вынужденного упрощения* на самом деле плохо различного механизма, то следует усовершенствовать эти допущения, что приведет, естественно, и к изменению модели.

В рассмотренном примере интерпретация временного рассогласования модели и действительности относилась как раз к первому типу. Проведенные сопоставления модельных и экспериментальных данных по распределению заработной платы работников за ряд лет (1956–1972 гг., см. [Айвазян С. А., 1980]) четко обозначили период их резкого рассогласования (1960–1966 гг.). Однако по мере удаления этого периода в прошлое прослеживается явная тенденция к сближению модельных и реальных данных. Более внимательный анализ показал, что момент резкого рассогласования

следовал непосредственно за весьма существенным директивным вмешательством в существующие тарифные условия, вмешательством, которое, как показал дальнейший ход развития, плохо согласовывалось с целым рядом объективных экономических закономерностей. И факт, что в дальнейшем мы наблюдаем сближение модельных и реальных данных, говорит лишь о том, что эти объективные экономические закономерности постепенно все более сказывались на характере распределения, все более «выступали на поверхность», отвоевывая себе те или иные правовые формы.

ВЫВОДЫ

1. Предлагаемый учебник посвящен описанию математико-статистического инструментария экономических исследований, который охватывает как традиционные методы эконометрики (регрессионный анализ, анализ временных рядов и системы одновременных уравнений), так и ряд наиболее актуальных, в смысле социально-экономических приложений, разделов прикладной статистики (кластер-анализ и расщепление смесей распределений, дискриминантный анализ, методы снижения размерности, построение интегральных индикаторов и др.).

2. Изложение в учебнике строится в соответствии с современными университетскими учебными планами подготовки экономистов и экономистов-статистиков и охватывает весь блок математико-статистического инструментария, в который входят следующие дисциплины: дескриптивная статистика, теория вероятностей, математическая статистика, многомерный статистический анализ, эконометрика.

3. *Теория вероятностей* — математическая наука, предназначенная для разработки (и исследования свойств) математических моделей, имитирующих механизмы функционирования реальных явлений или систем, условия «жизни» которых включают в себя неизбежность «мешающего» влияния большого числа случайных (т. е. не поддающихся строгому учету и контролю) факторов.

4. *Математическая статистика* — система основанных на теоретико-вероятностных моделях понятий, приемов и математических методов, предназначенных для сбора, систематизации, интерпретации и обработки статистических данных с целью получения научных и практических выводов. Одно из главных назначений методов математической статистики — обоснованный выбор среди множества возможных теоретико-вероятностных моделей той модели, которая наилучшим образом соответствует имеющимся в распоряжении исследователя статистическим данным, характеризующим реальное поведение конкретной исследуемой си-

стемы.

5. *Прикладная статистика* — научная дисциплина, разрабатывающая и систематизирующая понятия, приемы, математические методы и модели, предназначенные для организации сбора, стандартной записи, систематизации и обработки статистических данных с целью их удобного представления, интерпретации и получения научных и практических выводов.

6. *Эконометрика* — экономико-математическая научная дисциплина, разрабатывающая и использующая методы, модели, приемы, позволяющие придавать конкретное количественное выражение общим (качественным) закономерностям экономической теории на базе экономической статистики и с использованием математико-статистического инструментария.

7. Теория вероятностей и математическая статистика являются по отношению к прикладной статистике и эконометрике разработчиками и поставщиками существенной части используемого в них математического аппарата. Однако доводка и развитие этого аппарата, подчиненные требованиям и запросам различного типа приложений, производятся уже в рамках дисциплин «прикладная статистика» и «эконометрика».

8. *Границы применимости вероятностно-статистических методов* определяются, строго говоря, требованием соблюдения (хотя бы приблизительно) в исследуемой реальной действительности условий статистического ансамбля, а именно: а) возможностью (хотя бы мысленно реально представимой) многократного повторения наших экспериментов или наблюдений в одних и тех же условиях; б) наличием большого числа случайных факторов, характеризующих условия проведения наших экспериментов (наблюдений) и не позволяющих делать полностью предопределенного (детерминированного) заключения о том, произойдет или не произойдет в результате этих экспериментов интересующее нас событие.

9. *Строгит математических методов, позволяющих точно определять, находимся ли мы в условиях статистического ансамбля, не существует*: любая вероятностная модель, так же как и любая математическая модель вообще, есть лишь некоторое приближение к исследуемой реальной действительности. Можно лишь условно разделять исследуемые реальные ситуации на: 1) относящиеся к области высокой работоспособности вероятностно-статистических методов; 2) лишь с натяжкой укладывающиеся в рамки статистического ансамбля (категория допустимых вероятностно-статистических приложений); 3) недопустимые для вероятностно-статистических приложений. Однако и применительно к последней категории ситуаций в ряде случаев помимо методов анали-

за данных используют статистические методы, основанные на концепции субъективных вероятностей.

10. Всякая математическая модель является упрощенным представлением действительности, и искусство ее построения состоит в том, чтобы совместить как можно большую лаконичность параметризации модели с достаточной адекватностью описания изучаемой действительности или, другими словами, чтобы достигнуть максимальной концентрации реальности в простой математической форме.

11. *Вероятностная модель* — это математическая модель, имитирующая механизм функционирования гипотетического (не конкретного) реального явления или системы стохастической природы; *вероятностно-статистическая модель* — это вероятностная модель, значения отдельных характеристик (параметров) которой оцениваются по результатам наблюдений, характеризующим функционирование моделируемого *конкретного* (а не гипотетического) явления или системы; вероятностно-статистическая модель, описывающая механизм функционирования экономической или социально-экономической системы, называется *эконометрической*.

12. Важнейшим условием достижения высокой «работоспособности» модели является успешная реализация второго этапа моделирования, т. е. проведение тщательного предмодельного анализа физической сущности изучаемого явления с целью формирования добротной априорной информации и ее использования при выводе (или выборе) общего вида искомой модели. *Вынужденная (но нежелательная) альтернатива к такому подходу это логика «черного ящика»*, т. е. чисто формальная аппроксимация реальных данных.

Основы Теории Вероятностей

Глава 1. Правила действий со случайными событиями и вероятностями их осуществления

Глава 2. Случайные величины (исследуемые признаки)

Глава 3. Модели законов распределения вероятностей, наиболее распространенные в практике статистических исследований

Глава 4. Основные результаты теории вероятностей

Глава 5. Цепи Маркова

ГЛАВА 1. ПРАВИЛА ДЕЙСТВИЙ СО СЛУЧАЙНЫМИ СОБЫТИЯМИ И ВЕРОЯТНОСТЯМИ ИХ ОСУЩЕСТВЛЕНИЯ

1.1. Дискретное вероятностное пространство

1.1.1. Процесс регистрации наблюдения на объекте исследуемой совокупности (случайный эксперимент)

При построении любой математической теории прежде всего следует договориться об *определениях* некоторых понятий и о принятии в качестве исходных фактов, не требующих доказательства, некоторых допущений (*аксиом*). Отправным понятием теории вероятностей является понятие *случайного эксперимента*, который определяется как процесс регистрации наблюдения на единице обследуемой совокупности, произведенный в *условиях статистического ансамбля*. В зависимости от конкретного содержания этого понятия и применяется соответствующий математический аппарат к решению той или иной конкретной задачи. *Одна и та же реальная задача допускает в зависимости от конечных целей исследования несколько вариантов конкретизации понятия «случайный эксперимент»*. Так, при контроле изделия по альтернативному признаку (т. е. когда в результате контроля одного изделия оно признается либо годным, либо дефектным) пользуются в зависимости от конечных целей исследования по меньшей мере тремя вариантами интерпретации этого понятия: а) контроль одного изделия; б) контроль партии (выборки), состоящей из N изделий; в) контроль двух партий (выборок), состоящих соответственно из N_1 и N_2 изделий (схема повторной выборки). Очевидно, в первом варианте могут быть только два различных исхода («годное – дефектное»), во втором, если нас интересует только общее количество обнаруженных в

партии дефектных изделий и не интересует порядок, в котором они были обнаружены, — $(N + 1)$ возможных исходов (в партии обнаружено нуль, или одно, или два, ..., или N дефектных изделий), а в третьем — при том же подходе — $(N_1 + 1)(N_2 + 1)$ возможных исходов (результат каждого случайного эксперимента задается в этом случае парой чисел (ξ, η) , где ξ и η — числа дефектных изделий, обнаруженных соответственно в первой и во второй партиях).

Точно так же при анализе результатов выбрасывания игральных костей можно понимать под случайным экспериментом одно бросание (соответственно будем иметь шесть возможных исходов), а можно — последовательность из заданного числа бросаний, как мы и должны были бы поступить при строгой формализации примера игры с четырехкратным бросанием игральной кости, рассмотренного в п. В.2 (нетрудно подсчитать, что общее число возможных исходов в этом примере выразится числом $6^4 = 1296$).

Число возможных исходов случайного эксперимента не всегда можно пересчитать. Всякий случайный эксперимент, связанный с необходимостью фиксации величины какого-либо параметра обследуемого объекта, измеряемого в физических единицах *непрерывной природы* (температура, давление, время, вес, размеры и т. п.), имеет, как принято говорить, *континуальное множество* возможных исходов¹. Однако о том, как был осуществлен перенос построения строгой вероятностной теории на этот общий случай, речь будет идти в следующем пункте.

1.1.2. Случайные события и правила действий с ними

Как уже упоминалось, с каждым случайным экспериментом связано понятие совокупности всех возможных его исходов. Каждый из этих возможных исходов будем называть *элементарным (неразложимым) событием* (или *элементарным исходом*), а совокупность всех таких возможных исходов — *пространством элементарных событий* (исходов). Таким образом, в результате анализируемого случайного эксперимента обязательно происходит одно из элементарных событий, причем одновременно с ним не может произойти ни одно из остальных элементарных событий (о событиях, обладающих последним свойством, говорят, что они являются

¹ Как известно, множество элементов, составляющих какую-либо совокупность, может быть *конечным* (если число элементов в нем конечно), *счетным* (если его элементы можно занумеровать числами $1, 2, \dots, n, \dots$: множество целых чисел, множество рациональных чисел и т. п.) и *континуальным* (множество всех точек числовой прямой, плоскости и т. п.).

несовместными, см. ниже).

Рассмотрим пока лишь *дискретный случай*, т. е. только те ситуации, когда все элементарные события можно занумеровать числами $1, 2, \dots, n, \dots$. Иначе говоря, рассматриваемое пространство элементарных событий состоит лишь из конечного или счетного множества элементарных событий (обозначим их $\omega_1, \omega_2, \dots, \omega_n, \dots$), а сам факт «пространство Ω состоит из элементарных событий $\omega_1, \omega_2, \dots, \omega_n, \dots$ » кратко обозначим

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_n, \dots\},$$

или

$$\Omega = \{\omega_i\}, \quad i = 1, 2, \dots$$

(1.1)

Рассмотрим несколько примеров случайных экспериментов и соответствующих им пространств элементарных событий.

Пример 1.1. Подбрасывание монеты:

$$\Omega = \{\omega_1 = \text{аверс (лицевая сторона монеты)}; \\ \omega_2 = \text{реверс (обратная сторона монеты)}\}.$$

Пример 1.2. Выбрасывание одной игральной кости:

$$\Omega = \{\omega_1 = 1; \omega_2 = 2; \omega_3 = 3; \omega_4 = 4; \omega_5 = 5; \omega_6 = 6\}.$$

Пример 1.3. Четырехкратное бросание игральной кости:

$$\Omega = \{\omega_1 = (1, 1, 1, 1); \omega_2 = (1, 1, 1, 2); \dots; \omega_{1296} = (6, 6, 6, 6)\}.$$

Пример 1.4. Проверка (по альтернативному признаку) одного изделия, случайно отобранного из продукции массового производства:

$$\Omega = \{\omega_1 = \text{годно}; \omega_2 = \text{дефектно}\}.$$

Пример 1.5. Проверка (по альтернативному признаку) N изделий, случайно отобранных из продукции массового производства:

$$\Omega = \{\omega_1 = 0; \omega_2 = 1; \dots; \omega_{N+1} = N\},$$

где ω — число обнаруженных дефектных изделий.

Пример 1.6. Проверка (по альтернативному признаку) двух выборок, состоящих соответственно из N_1 и N_2 изделий, случайно отобранных из продукции массового производства:

$$\Omega = \{\omega_1 = (0, 0); \omega_2 = (0, 1); \dots; \omega_{(N_1+1)(N_2+1)} = (N_1, N_2)\}.$$

Пример 1.7. Определение ω — числа сбоев станков автоматической линии за наугад выбранную смену:

$$\Omega = \{\omega_1 = 0; \omega_2 = 1; \dots; \omega_n = n - 1; \omega_{n+1} = n; \dots\}.$$

Кроме *элементарных* событий исследователю приходится иметь дело с так называемыми *составными* (или *разложимыми*) событиями.

Событие C называется составным (разложимым), если можно указать по меньшей мере два таких элементарных события ω_{i_1} и ω_{i_2} , что из осуществления каждого из них в отдельности следует факт осуществления события C . Этот факт коротко формулируют — «событие C состоит из элементарных событий ω_{i_1} и ω_{i_2} » — и записывают символически в виде $C = \{\omega_{i_1}, \omega_{i_2}\}$.

Пользуясь такой терминологией, *случайным событием A называют любое подмножество $\{\omega_{i_1}, \omega_{i_2}, \dots, \omega_{i_k}, \dots\}$ пространства элементарных событий (1.1), т. е.*

$$A = \{\omega_{i_1}, \omega_{i_2}, \dots, \omega_{i_k} \dots\},$$

что следует понимать так: осуществление любого из элементарных событий $\omega_{i_1}, \dots, \omega_{i_k}, \dots$, «входящих в A », влечет за собой осуществление события A .

Поясним эту терминологию на некоторых из примеров 1.1–1.7.

В примере 1.2 события $A_1 = \{\text{выпадет четное число очков}\}$ и $A_2 = \{\text{выпавшее число очков не превзойдет 3}\}$ запишутся соответственно $A_1 = \{\omega_2, \omega_4, \omega_6\}$ и $A_2 = \{\omega_1, \omega_2, \omega_3\}$.

В примере 1.3 интересовавшее нас в п. В.2 событие $A = \{\text{хотя бы один раз при четырехкратном бросании кости появится шестерка}\}$ будет состоять из всех тех четырехмерных векторов ω_i , в которых хотя бы одна из компонент равна шести (можно подсчитать, что число таких векторов равно 671).

В примере 1.5 событие $A = \{\text{число обнаруженных дефектных изделий не более 4}\}$, очевидно, может быть записано $A = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$.

В примере 1.7 событие $A = \{\text{число зафиксированных за смену сбоев станков автоматической линии меньше 3}\}$ можно представить в виде $A = \{\omega_1, \omega_2, \omega_3\}$.

Из определения теории вероятностей (см. п. В.2) следует, что в первую очередь надо запастись определенной структурой всех возможных связей между случайными событиями или, другими словами, *определенными правилами действий с событиями* и соответствующей терминологией.

Сумма (объединение) событий $A_1, A_2, \dots, A_k$ — это такое событие A ($A = A_1 + A_2 + \dots + A_k$), которое заключается в наступлении хотя бы одного из событий $A_1, A_2, \dots, A_k$. На языке элементарных событий,

следовательно, сумма событий $A_1, A_2, \dots, A_k$ определяется как событие A , состоящее из *всех различных* элементарных событий, составляющих события $A_1, A_2, \dots, A_k$. Таким образом, если нас интересует, например, сумма A событий $A_1 = \{\text{число обнаруженных дефектных изделий не более 4}\} = \{\omega_1, \omega_2, \dots, \omega_5\}$ (см. пример 1.5) и $A_2 = \{\text{число обнаруженных дефектных изделий заключено между 2 и 6 включительно}\} = \{\omega_3, \omega_4, \omega_5, \omega_6, \omega_7\}$, то, очевидно, $A = A_1 + A_2 = \{\omega_1, \omega_2, \dots, \omega_5, \omega_6, \omega_7\}$, так как о том, что событие A произошло, будет сигнализировать реализация любого из элементарных событий $\omega_1, \dots, \omega_7$.

Произведение (пересечение) событий $A_1, A_2, \dots, A_k$ — это такое событие A ($A = A_1 \cdot A_2 \cdot \dots \cdot A_k$), которое заключается в обязательном наступлении всех событий $A_1, A_2, \dots, A_k$. На языке элементарных событий, следовательно, произведение событий $A_1, A_2, \dots, A_k$ определяется как событие A , состоящее лишь из тех элементарных событий, которые *одновременно входят* во все рассматриваемые события $A_1, A_2, \dots, A_k$. Так что произведением упомянутых выше событий $A_1 = \{\omega_1, \omega_2, \dots, \omega_5\}$ и $A_2 = \{\omega_3, \omega_4, \dots, \omega_7\}$ (см. пример 1.5) будет, очевидно, событие $A = A_1 \cdot A_2 = \{\omega_3, \omega_4, \omega_5\}$, так как реализация каждого из элементарных событий ω_3, ω_4 и ω_5 в отдельности означает, как легко видеть, одновременное наступление событий A_1 и A_2 .

Разность событий A_1 и A_2 — это такое событие A , которое заключается в одновременном осуществлении двух фактов: событие A_1 произошло, а событие A_2 не произошло. На языке элементарных событий, следовательно, разность $A = A_1 - A_2$ определяется как событие, состоящее из всех тех элементарных событий, которые входят в A_1 , но не входят в A_2 . Так что разностью рассмотренных выше в схеме примера 1.5 событий $A_1 = \{\omega_1, \omega_2, \dots, \omega_5\}$ и $A_2 = \{\omega_3, \omega_4, \dots, \omega_7\}$ будет, очевидно, событие $A = A_1 - A_2 = \{\omega_1, \omega_2\}$, т. е. событие, заключающееся в том, что число обнаруженных в партии дефектных изделий окажется не превосходящим единицы.

Противоположное (дополнительное) событие (к событию A) — это такое событие $\bar{A}$, которое состоит в ненаступлении события A . На языке элементарных событий, следовательно, $\bar{A}$ определяется как событие, состоящее из всех возможных элементарных событий, не входящих в A . Поэтому, используя понятие разности событий, можно записать $\bar{A} = \Omega - A$. Так, например, событием, противоположным событию $A_1 = \{\omega_1, \omega_2, \dots, \omega_5\}$ из примера 1.5, очевидно, будет событие, заключающееся в том, что число обнаруженных в партии объема N дефектных изделий окажется большим 4, т. е. $\bar{A}_1 = \{\omega_6, \omega_7, \dots, \omega_{N+1}\}$.

Достоверное событие определяется как событие, состоящее из всех возможных элементарных событий, т.е. это событие $\Omega = \{\omega_i\}$, $i = 1, 2, \dots$ (см. (1.1)). Название этого события оправдано тем обстоятельством, что пространство элементарных событий состоит из *всех* возможных исходов, т. е. в результате анализируемого случайного эксперимента *обязательно* произойдет одно из элементарных событий ω_i , а следовательно, тот факт, что событие Ω произойдет, *достоверен*.

Невозможное (пустое) событие $\emptyset$ — это событие, противоположное достоверному, т.е. $\emptyset = \Omega - \Omega = \bar{\Omega}$. Из определения непосредственно следует объяснение названия этого события: *оно не содержит ни одного элементарного события* ω_i и, следовательно, при реализации исследуемого случайного эксперимента его осуществление невозможно.

Несовместными называются события $A_1, A_2, \dots, A_k$, если в результате исследуемого случайного эксперимента никакие два из них не могут произойти одновременно. На языке элементарных событий это означает, что среди событий $A_1, A_2, \dots, A_k$ нельзя найти такую пару событий A_i и A_j , в которой обнаружилось бы хотя бы по одному общему элементарному событию. Используя понятия произведения событий и невозможного события, можно определить несовместные события как такую последовательность событий $A_1, A_2, \dots, A_k$, для любой пары которой A_i и A_j справедливо соотношение $A_i \cdot A_j = \emptyset$. Очевидно, любые пары $(A, \bar{A})$ противоположных событий являются несовместными. Таковыми же являются (по определению) и все элементарные события.

Полная система событий — это такой набор несовместных событий $A_1, A_2, \dots, A_k$, который в сумме исчерпывает все пространство элементарных событий, т.е.

$$\begin{cases} A_1 + A_2 + \dots + A_k = \Omega; \\ A_i \cdot A_j = \emptyset \end{cases}$$

для всех $i, j = 1, 2, \dots, k$ и $i \neq j$.

Очевидно, набор всех элементарных событий можно рассматривать как частный случай полной системы событий. И вообще любое разбиение множества элементарных событий Ω на непересекающиеся классы дает полную систему событий, в которой каждое из событий задается соответствующим классом разбиения. Так, обратившись вновь к примеру 1.5, положив в нем для определенности $N = 100$ и обозначив буквой d число дефектных изделий, обнаруженных в партии, состоящей из 100 наугад извлеченных изделий, отобранных от стационарно действующего массового

производства, определим систему событий следующим образом:

$$A_1 = \{d = 0\} = \{\omega_1\}$$

— партия отличного качества;

$$A_2 = \{1 \leq d \leq 5\} = \{\omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$$

— брак в пределах допустимой нормы (партия принимается);

$$A_3 = \{6 \leq d \leq 10\} = \{\omega_7, \omega_8, \dots, \omega_{11}\}$$

— брак выше допустимой нормы (партия принимается по сниженным расценкам);

$$A_4 = \{d \geq 11\} = \{\omega_{12}, \omega_{13} \dots, \omega_{101}\}$$

— партия целиком возвращается.

Очевидно, события A_1, A_2, A_3, A_4 образуют полную систему событий.

1.1.3. Вероятностное пространство. Вероятности и правила действий с ними

Для полного описания механизма исследуемого случайного эксперимента недостаточно задать лишь пространство элементарных событий. Очевидно, наряду с перечислением всех возможных исходов исследуемого случайного эксперимента мы должны также знать, как часто в длинной серии таких экспериментов могут происходить те или другие элементарные события. Действительно, возвращаясь, скажем, к примерам 1.1–1.7, легко представить себе, что в рамках каждого из описанных в них пространств элементарных событий можно рассмотреть бесчисленное множество случайных экспериментов, существенно различающихся по своему механизму. Так, в примерах 1.1–1.3 мы будем иметь существенно различающиеся относительные частоты появления одних и тех же элементарных исходов, если будем пользоваться различными монетами и игральными костями (симметричными, со слегка смещенным центром тяжести, с сильно смещенным центром тяжести и т. п.). В примерах 1.4–1.7 частота появления дефектных изделий, характер засоренности дефектными изделиями проконтролированных партий и частоты появления определенного числа сбоев станков автоматической линии будут зависеть от уровня технологической оснащенности изучаемого производства: при одном и том же пространстве элементарных событий частота появления «хороших» элементарных исходов будет выше в производстве с более высоким уровнем технологии.

Для построения (в дискретном случае) полной и законченной математической теории случайного эксперимента — **теории вероятностей** — помимо уже введенных исходных понятий *случайного эксперимента*, *элементарного исхода* и *случайного события* необходимо запастись еще *одним исходным допущением (аксиомой)*, постулирующим существование вероятностей элементарных событий (удовлетворяющих определенной нормировке), и *определением* вероятности любого случайного события.

Аксиома. Каждому элементу ω_i пространства элементарных событий Ω соответствует некоторая неотрицательная числовая характеристика p_i шансов его появления, называемая вероятностью события ω_i , причем

$$p_1 + p_2 + \dots + p_n + \dots = \sum_{i: \omega_i \in \Omega} p_i = 1 \quad (1.2)$$

(отсюда, в частности, следует, что $0 \leq p_i \leq 1$ для всех i).

Определение вероятности события. Вероятность любого события A определяется как сумма вероятностей всех элементарных событий, составляющих событие A , т.е. если использовать символику $P\{A\}$ для обозначения «вероятности события A », то

$$P\{A\} = \sum_{\omega_i \in A} P\{\omega_i\} = \sum_{i: \omega_i \in A} p_i. \quad (1.3)$$

Отсюда и из (1.2) непосредственно следует, что всегда $0 \leq P\{A\} \leq 1$, причем вероятность достоверного события равна единице, а вероятность невозможного события равна нулю. Все остальные понятия и правила действий с вероятностями и событиями будут уже производными от введенных выше четырех исходных определений (случайного эксперимента, элементарного исхода, случайного события и его вероятности) и одной аксиомы.

Таким образом, для исчерпывающего описания механизма исследуемого случайного эксперимента (в дискретном случае) необходимо задать конечное или счетное множество всех возможных элементарных исходов Ω и каждому элементарному исходу ω_i поставить в соответствие некоторую неотрицательную (не превосходящую единицы) числовую характеристику p_i , интерпретируемую как вероятность появления исхода ω_i (будем обозначать эту вероятность символами $P\{\omega_i\}$), причем установленное соответствие типа $\omega_i \leftrightarrow p_i$ должно удовлетворять требованию нормировки (1.2).

Вероятностное пространство как раз и является понятием, формализующим такое описание механизма случайного эксперимента. Задать

вероятностное пространство — это значит задать пространство элементарных событий Ω и определить в нем вышеуказанное соответствие типа

$$\omega_i \leftrightarrow p_i = P\{\omega_i\}. \quad (1.4)$$

Очевидно, соответствие типа (1.4) может быть задано различными способами: с помощью таблиц, графиков, аналитических формул, наконец, алгоритмически.

Как же построить вероятностное пространство, соответствующее исследуемому реальному комплексу условий? С наполнением конкретным содержанием понятий случайного эксперимента, элементарного события, пространства элементарных событий, а в дискретном случае — и любого разложимого случайного события, затруднений, как правило, не бывает. А вот определить из конкретных условий решаемой задачи вероятности $P\{\omega_i\}$ отдельных элементарных событий не так-то просто! С этой целью используется один из следующих трех подходов.

Априорный подход к вычислению вероятностей $P\{\omega_i\}$ заключается в теоретическом, умозрительном анализе специфических условий данного конкретного случайного эксперимента (до проведения самого эксперимента). В ряде ситуаций этот предопытный анализ позволяет теоретически обосновать способ определения искомых вероятностей. Например, возможен случай, когда пространство всех возможных элементарных исходов состоит из конечного числа N элементов, причем условия производства исследуемого случайного эксперимента таковы, что вероятности осуществления каждого из этих N элементарных исходов нам представляются равными (именно в такой ситуации мы находимся при подбрасывании симметричной монеты, бросании правильной игральной кости, случайном извлечении игральной карты из хорошо перемешанной колоды и т. п.). В силу аксиомы (1.2) вероятность каждого элементарного события равна в этом случае $1/N$. Это позволяет получить простой рецепт и для подсчета вероятности любого события: если событие A содержит N_A элементарных событий, то в соответствии с определением (1.3)

$$P\{A\} = \frac{N_A}{N}. \quad (1.3')$$

Смысл формулы (1.3') состоит в том, что вероятность события в данном классе ситуаций может быть определена как отношение числа благоприятных исходов (т. е. элементарных исходов, входящих в это событие) к числу всех возможных исходов (так называемое *классическое определение вероятности*). В современной трактовке формула (1.3') не является определением вероятности: она применима лишь в том частном случае, когда все элементарные исходы равновероятны.

Апостериорно-частотный подход к вычислению вероятностей $P\{\omega_i\}$ отталкивается, по существу, от определения вероятности, принятого так называемой частотной концепцией вероятности. В соответствии с этой концепцией вероятность $P\{\omega_i\}$ *определяется* как предел относительной частоты появления исхода ω_i в процессе неограниченного увеличения общего числа случайных экспериментов n , т. е.

$$p_i = P\{\omega_i\} = \lim_{n \rightarrow \infty} \frac{m_n(\omega_i)}{n}, \quad (1.5)$$

где $m_n(\omega_i)$ — число случайных экспериментов (из общего числа n произведенных случайных экспериментов), в которых зарегистрировано появление элементарного события ω_i . Соответственно для практического (приближенного) определения вероятностей p_i предлагается брать относительные частоты появления события ω_i в достаточно длинном ряду случайных экспериментов¹. Подобный способ *вычисления* вероятностей p_i не противоречит современной (аксиоматической) концепции теории вероятностей, поскольку последняя построена таким образом, что эмпирическим (или выборочным) аналогом объективно существующей вероятности $P\{A\}$ любого события A является относительная частота осуществления этого события в ряду из n независимых испытаний. Разными в этих двух концепциях оказываются *определения* вероятностей: в соответствии с частотной концепцией вероятность не является объективным, *существующим до опыта*, свойством изучаемого явления, а появляется *только в связи с проведением опыта* или наблюдения; это приводит к смешению теоретических (истинных, обусловленных реальным комплексом условий «существования» исследуемого явления) вероятностных характеристик и их эмпирических (выборочных) аналогов.

Как пишет Г. Крамер, «указанное определение вероятности можно сравнить, например, с определением геометрической точки как предела пятен мела неограниченно убывающих размеров, но подобного определения современная аксиоматическая геометрия не вводит» ([Крамер Г., с. 172]).

¹ Немецкий математик Р. Мизес, с именем которого связывают развитие частотной концепции вероятности, считал, что каждой вероятностной задаче обязательно соответствует некоторый реальный процесс (удовлетворяющий введенным им условиям «статистического коллектива»), а потому полагал теорию вероятностей дисциплиной естественно-научной («наукой о явлениях реального мира»), но не математической. Сущность понятия мизесского «коллектива» — в требовании *реальной* массовости изучаемого явления (и соответствующего эксперимента), существования пределов вида (1.5) и независимости этих пределов от того, по какой подпоследовательности произведенных случайных экспериментов, отобранной из всей такой последовательности, мы этот предел будем вычислять.

Мы не будем здесь останавливаться на математических изъясных частотной концепции вероятности. Отметим лишь принципиальные сложности реализации *вычислительного приема* получения приближенных значений p_i с помощью относительных частот $m_n(\omega_i)/n$. *Во-первых*, сохранение неизменными условий случайного эксперимента (т.е. сохранение условий статистического ансамбля), при котором оказывается справедливым допущение о тенденции относительных частот группироваться вокруг постоянного значения, не может поддерживаться неограниченно долго и с высокой точностью. Поэтому для оценки вероятностей p_i с помощью относительных частот $m_n(\omega_i)/n$ не имеет смысла брать слишком длинные ряды (т.е. слишком большие n) и потому же, кстати, точный переход к пределу (1.5) не может иметь реального смысла. *Во-вторых*, в ситуациях, когда мы имеем достаточно большое число возможных элементарных исходов (а они могут образовывать и бесконечное, и даже, как это было уже отмечено в п. 1.1.1, континуальное множество), даже в сколь угодно длинном ряду случайных экспериментов, мы будем иметь *возможные* исходы ω_i , ни разу не осуществившиеся в ходе нашего эксперимента; да и по другим редко реализующимся возможным исходам полученные с помощью относительных частот приближенные значения вероятностей будут в этих условиях крайне мало надежными.

Апостериорно-модельный подход к заданию вероятностей $P\{\omega_i\}$, отвечающему конкретно исследуемому реальному комплексу условий, является в настоящее время, пожалуй, наиболее распространенным и наиболее практически удобным. Логика этого подхода следующая. С одной стороны, в рамках априорного подхода, т.е. в рамках теоретического, умозрительного анализа возможных вариантов специфики гипотетичных реальных комплексов условий разработан и исследован набор *модельных вероятностных* пространств (биномиальное, пуассоновское, нормальное, показательное и т. п., см. гл. 3). С другой стороны, исследователь располагает *результатами ограниченного ряда случайных экспериментов*. Далее с помощью специальных математико-статистических приемов (основанных на методах статистического оценивания неизвестных параметров и статистической проверки гипотез, см. гл. 6–8) исследователь как бы прилагает гипотетичные модели вероятностных пространств к имеющимся у него результатам наблюдения (отражающим специфику изучаемой реальной действительности) и оставляет для дальнейшего использования лишь ту модель или те модели, которые не противоречат этим результатам и в некотором смысле наилучшим образом им соответствуют.

Опишем теперь основные правила действий с вероятностями событий, являющиеся следствиями принятых выше определений и аксиомы.

Вероятность суммы событий (теорема сложения вероятностей). Сформулируем и докажем правило вычисления вероятности суммы двух событий A_1 и A_2 . Для этого разобьем каждое из множеств элементарных событий, составляющих события A_1 и A_2 , на две части:

$$A_1 = A_1^0 + A_{12}; \quad A_2 = A_2^0 + A_{12},$$

где A_k^0 объединяет все элементарные события ω_i , входящие в A_k , но не входящие в A_m ($k, m = 1, 2; k \neq m$), а A_{12} состоит из всех тех элементарных событий, которые одновременно входят и в A_1 , и в A_2 . Пользуясь определением (1.3) и определением произведения событий A_1 и A_2 , имеем:

$$P\{A_1\} = \sum_{i: \omega_i \in A_1} p_i = \sum_{i: \omega_i \in A_1^0} p_i + \sum_{i: \omega_i \in A_{12}} p_i = P\{A_1^0\} + P\{A_1 \cdot A_2\}; \quad (1.6)$$

$$P\{A_2\} = \sum_{i: \omega_i \in A_2} p_i = \sum_{i: \omega_i \in A_2^0} p_i + \sum_{i: \omega_i \in A_{12}} p_i = P\{A_2^0\} + P\{A_1 \cdot A_2\}. \quad (1.7)$$

В то же время в соответствии с определением суммы событий $A = A_1 + A_2$ и с (1.3) имеем

$$P\{A_1 + A_2\} = \sum_{i: \omega_i \in A_1^0} p_i + \sum_{i: \omega_i \in A_2^0} p_i + \sum_{i: \omega_i \in A_{12}} p_i. \quad (1.8)$$

Из (1.6), (1.7) и (1.8) получаем *формулу сложения вероятностей* (для двух событий):

$$P\{A_1 + A_2\} = P\{A_1\} + P\{A_2\} - P\{A_1 A_2\}. \quad (1.9)$$

Формула (1.9) сложения вероятностей может быть обобщена на случай произвольного числа слагаемых:

$$P\{A_1 + A_2 + \dots + A_k\} = \sum_{i=1}^k P\{A_i\} - \Delta_1 + \Delta_2 - \dots + (-1)^{k-1} \Delta_{k-1}, \quad (1.9')$$

где «добавки» Δ_m ($m = 1, 2, \dots, k-1$) вычисляются в форме суммы вероятностей вида

$$\Delta_m = \sum_{i_1=1}^k \sum_{i_2=1}^k \dots \sum_{i_{m+1}=1}^k P\{A_{i_1} \cdot A_{i_2} \cdot \dots \cdot A_{i_{m+1}}\},$$

причем суммирование в правой части производится, очевидно, при условии, что все $i_1, i_2, \dots, i_{m+1}$ различны и $i_1 < i_2 < \dots < i_{m+1}$. В частном

случае, когда интересующая нас система $A_1, A_2, \dots, A_k$ состоит лишь из *несовместных* событий, все произведения вида $A_{i_1} \cdot A_{i_2} \cdot \dots \cdot A_{i_m+1}$ ($m \geq 1$) будут пустыми (или невозможными) событиями и соответственно формула (1.9') дает

$$P\{A_1 + A_2 + \dots + A_k\} = P\{A_1\} + P\{A_2\} + \dots + P\{A_k\}. \quad (1.9'')$$

Вероятность произведения событий (теорема умножения вероятностей). **Условная вероятность.** Рассмотрим ситуации, когда заранее поставленное условие или фиксация некоторого уже осуществившегося события исключают из числа возможных часть элементарных событий анализируемого вероятностного пространства. Так, анализируя совокупность из N изделий массового производства, содержащую N_1 изделий первого, N_2 — второго, N_3 — третьего и N_4 — четвертого сорта ($N_1 + N_2 + N_3 + N_4 = N$), мы рассматриваем вероятностное пространство с элементарными исходами $\omega_1, \omega_2, \omega_3$ и ω_4 и их вероятностями — соответственно $p_1 = N_1/N$, $p_2 = N_2/N$, $p_3 = N_3/N$ и $p_4 = N_4/N$ (здесь ω_i означает событие, заключающееся в том, что наугад извлеченное из совокупности изделие оказалось i -го сорта). Предположим, условия сортировки изделий таковы, что на каком-то этапе изделия первого сорта отделяются от общей совокупности и все вероятностные выводы (и, в частности, подсчет вероятностей различных событий) нам предстоит строить применительно к *урезанной* совокупности, состоящей только из изделий второго, третьего и четвертого сорта. В таких случаях принято говорить об *условных вероятностях*, т. е. о вероятностях, вычисленных при условии уже осуществленного некоторого события. В данном случае таким осуществленным событием является событие $B = \{\omega_2, \omega_3, \omega_4\}$, т. е. событие, заключающееся в том, что любое наугад извлеченное изделие является либо второго, либо третьего, либо четвертого сорта. Поэтому, если нас интересует подсчет условной вероятности события A (при условии, что событие B уже имеет место), заключающегося, например, в том, что наугад извлеченное изделие окажется второго или третьего сорта, то, очевидно, эта условная вероятность (обозначим ее $P\{A | B\}$) может быть определена следующим соотношением:

$$P\{A | B\} = \frac{N_2 + N_3}{N_2 + N_3 + N_4} = \frac{N_2 + N_3}{N} \bigg/ \frac{N_2 + N_3 + N_4}{N} = \frac{P\{AB\}}{P\{B\}}.$$

Как легко понять из этого примера, подсчет условных вероятностей — это, по существу, *переход в другое, урезанное заданным условием B пространство элементарных событий*, когда соотношение вероятностей элементарных событий в урезанном пространстве остается тем же,

что и в исходном (более широком), но все они нормируются (делятся на $P\{B\}$) для того, чтобы и в новом вероятностном пространстве выполнялось требование нормировки (1.2). Конечно, можно было бы не вводить терминологии с условными вероятностями, а просто использовать аппарат обычных («безусловных») вероятностей в новом пространстве. Запись в терминах вероятностей «старого» пространства бывает полезной в тех случаях, когда по условиям конкретной задачи мы должны все время помнить о существовании исходного, более широкого пространства элементарных событий.

Получим формулу условной вероятности в общем случае. Пусть B — событие (непустое), считающееся уже состоявшимся («условие»), а A — событие, условную вероятность которого $P\{A | B\}$ требует вычислить. Новое (урезанное) пространство элементарных событий $\tilde{\Omega}$ состоит только из элементарных событий, входящих в B , и, следовательно, их вероятности (с условием нормировки (1.2)) определяются соотношениями¹:

$$\tilde{p}_i = P_{\tilde{\Omega}}\{\omega_i\} = \begin{cases} \frac{P_{\Omega}\{\omega_i\}}{P_{\Omega}\{B\}} = \frac{p_i}{P_{\Omega}\{B\}}, & \text{если } \omega_i \in B; \\ 0, & \text{если } \omega_i \notin B. \end{cases} \quad (1.10)$$

По определению вероятность $P\{A | B\}$ — это вероятность события A в «урезанном» вероятностном пространстве $\{\tilde{\Omega}, \tilde{p}\}$, и, следовательно, в соответствии с (1.3) и (1.10)

$$\begin{aligned} P_{\Omega}\{A | B\} &= P_{\tilde{\Omega}}\{A\} = \sum_{\omega_i \in A} P_{\tilde{\Omega}}\{\omega_i\} = \\ &= \sum_{i: \omega_i \in AB} \left(\frac{p_i}{P_{\Omega}\{B\}} \right) = \frac{\sum_{i: \omega_i \in AB} p_i}{P_{\Omega}\{B\}}, \end{aligned}$$

т. е.

$$P_{\Omega}\{A | B\} = \frac{P_{\Omega}\{AB\}}{P_{\Omega}\{B\}}, \quad (1.11)$$

или, что то же,

$$P_{\Omega}\{AB\} = P_{\Omega}\{A | B\} \cdot P_{\Omega}\{B\}. \quad (1.11')$$

¹ Нижний индекс у буквы P , означающей «вероятность» (P — первая буква латинской транскрипции этого слова), поясняет, в каком именно вероятностном пространстве или пространстве элементарных событий производится вычисление вероятности соответствующего события.

Эквивалентные формулы (1.11) и (1.11') принято называть соответственно формулой *условной вероятности* и *правилом умножения вероятностей*.

Еще раз подчеркнем, что рассмотрение условных вероятностей различных событий *при одном и том же условии* B равносильно рассмотрению обычных вероятностей в другом (урезанном) пространстве элементарных событий $\tilde{\Omega} = B$ с пересчетом соответствующих вероятностей элементарных событий по формуле (1.10). Поэтому *все общие теоремы и правила действий с вероятностями остаются в силе и для условных вероятностей, если эти условные вероятности берутся при одном и том же условии*.

Независимость событий. Два события A и B называют независимыми, если

$$P\{AB\} = P\{A\} \cdot P\{B\}. \quad (1.12)$$

Для пояснения естественности такого определения вернемся к теореме умножения вероятностей (1.11') и посмотрим, в каких ситуациях из нее следует (1.12). Очевидно, это может быть тогда, когда условная вероятность $P\{A \mid B\}$ равна соответствующей безусловной вероятности $P\{A\}$, т. е., грубо говоря, тогда, когда знание того, что произошло событие B , никак не влияет на оценку шансов появления события A .

Распространение определения независимости на систему более чем двух событий выглядит следующим образом. События $A_1, A_2, \dots, A_k$ называются *взаимно (или совместно) независимыми*, если для любых пар, троек, четверок и т. д. событий, отобранных от этого набора событий, справедливы следующие правила умножения:

$$\begin{aligned} P\{A_{i_1} A_{i_2}\} &= P\{A_{i_1}\} P\{A_{i_2}\}; \\ P\{A_{j_1} A_{j_2} A_{j_3}\} &= P\{A_{j_1}\} P\{A_{j_2}\} P\{A_{j_3}\}; \\ &\dots\dots\dots \\ P\{A_1 \cdot A_2 \cdots A_k\} &= P\{A_1\} P\{A_2\} \cdots P\{A_k\}. \end{aligned} \quad (1.13)$$

Очевидно, в первой строке подразумевается

$$C_k^2 = \frac{k(k-1)}{1 \cdot 2}$$

(число сочетаний из k по два) уравнений, во второй — C_k^3 уравнений и т. д. Всего, следовательно, (1.13) объединяет $C_k^2 + C_k^3 + \cdots + C_k^k = 2^k - k - 1$ условий. В то же время C_k^2 условий первой строки достаточно для обеспечения *попарной независимости* этих событий. И хотя попарная и

взаимная независимость системы событий, строго говоря, не одно и то же, их различие представляет скорее теоретический, чем практический интерес.

Свойство независимости событий сильно облегчает анализ различных вероятностей, связанных с исследуемой системой событий. Достаточно сказать, что если в общем случае для описания вероятностей всевозможных комбинаций событий системы $A_1, A_2, \dots, A_k$ нужно задать 2^k вероятностей, то в случае взаимной независимости этих событий достаточно лишь k вероятностей $P\{A_1\}, P\{A_2\}, \dots, P\{A_k\}$.

Независимые события весьма часто встречаются в изучаемой реальной действительности: они осуществляются в экспериментах (наблюдениях), проводимых независимо друг от друга в обычном физическом смысле.

Именно свойство независимости исходов четырех последовательных бросаний игральной кости позволило (с помощью (1.13)) легко подсчитать вероятность невыпадения (ни при одном из этих бросаний) шестерки в задаче из п. В.2. Действительно, обозначив A_i событие, заключающееся в невыпадении шестерки в i -м бросании ($i = 1, 2, 3, 4$), и учитывая, что $P\{A_i\} = 5/6$ для всех $i = 1, 2, 3, 4$, получаем

$$\begin{aligned} P\{A_1 A_2 A_3 A_4\} &= P\{A_1\}P\{A_2\}P\{A_3\}P\{A_4\} \\ &= (5/6)^4 = 625/1296. \end{aligned}$$

Формула полной вероятности. При решении многих практических задач зачастую сталкиваются с ситуацией, когда прямое вычисление вероятности интересующего нас события A трудно или невозможно, в то время как вполне доступно вычисление (или задание) условных вероятностей того же события (при различных условиях). В случае, когда условия $B_1, B_2, \dots, B_k$, при которых известны (или легко вычисляемы) условные вероятности события A , образуют полную систему событий (см. п. 1.1.2), для подсчета вероятности $P\{A\}$ можно использовать соотношение

$$\begin{aligned} P\{A\} &= P\{A \mid B_1\}P\{B_1\} + \\ &+ P\{A \mid B_2\}P\{B_2\} + \dots + P\{A \mid B_k\}P\{B_k\}, \end{aligned} \quad (1.14)$$

которое принято называть *формулой полной вероятности*.

Для доказательства формулы (1.14) заметим, что элементарные события, составляющие событие A , можно разбить на k непересекающихся групп, каждая из которых является общей частью (пересечением) события A с одним из событий B_i (эта возможность непосредственно вытекает из того, что события $B_1, B_2, \dots, B_k$ исчерпывают в сумме все пространство

элементарных событий и попарно не пересекаются), т. е.

$$A = AB_1 + AB_2 + \dots + AB_k.$$

Далее, воспользовавшись теоремой сложения вероятностей (1.9'') (применительно к несовместным событиям, каковыми являются события $AB_1, AB_2, \dots, AB_k$) и вычислив вероятность каждого из произведений AB_i по формуле произведения вероятностей (1.11'), мы и получаем (1.14).

Формула Байеса. Обратимся вначале к следующей задаче. На складе имеются приборы, изготовленные тремя заводами: 20% приборов на складе изготовлены заводом № 1, 50% — заводом № 2 и 30% — заводом № 3. Вероятности того, что в течение гарантийного срока прибору потребуется ремонт, для продукции каждого из заводов равны соответственно 0,2; 0,1; 0,3. Взятый со склада прибор не имел заводской маркировки и потребовал ремонта (в течение гарантийного срока). Каким заводом вероятнее всего был изготовлен прибор? Какова эта вероятность? Если обозначить A_i событие, заключающееся в том, что случайно взятый со склада прибор оказался изготовленным на i -м заводе ($i = 1, 2, 3$), а B — событие, заключающееся в том, что наугад отобранный от продукции всех трех заводов прибор оказался дефектным (потребовал ремонта), то сформулированная выше задача, очевидно, сводится к вычислению условных вероятностей $P\{A_i | B\}$ по заданным вероятностям $P\{A_1\} = 0,2$; $P\{A_2\} = 0,5$; $P\{A_3\} = 0,3$; $P\{B | A_1\} = 0,2$; $P\{B | A_2\} = 0,1$ и $P\{B | A_3\} = 0,3$. Поскольку события A_1, A_2, A_3 образуют полную систему, воспользуемся для выражения искомых вероятностей $P\{A_i | B\}$ известными нам основными правилами действий с вероятностями. По формуле условной вероятности (1.11)

$$P\{A_i | B\} = \frac{P\{A_i B\}}{P\{B\}}. \quad (1.15)$$

Числитель этой дроби по теореме умножения вероятностей (1.11') может быть представлен в виде

$$P\{A_i B\} = P\{BA_i\} = P\{B | A_i\}P\{A_i\}, \quad (1.16)$$

а знаменатель $P\{B\}$ выражается с помощью формулы полной вероятности (1.14):

$$P\{B\} = P\{B | A_1\}P\{A_1\} + P\{B | A_2\}P\{A_2\} + P\{B | A_3\}P\{A_3\}. \quad (1.17)$$

Подставляя (1.16) и (1.17) в (1.15), получаем

$$P\{A_i | B\} = \frac{P\{A_i\}P\{B | A_i\}}{\sum_{j=1}^3 P\{B | A_j\}P\{A_j\}}. \quad (1.18)$$

Воспользовавшись этой формулой, нетрудно подсчитать искомые вероятности:

$$P\{A_1 | B\} = \frac{0,2 \cdot 0,2}{0,2 \cdot 0,2 + 0,1 \cdot 0,5 + 0,3 \cdot 0,3} = \frac{0,04}{0,18} = 0,222;$$

$$P\{A_2 | B\} = \frac{0,5 \cdot 0,1}{0,18} = \frac{0,05}{0,18} = 0,278;$$

$$P\{A_3 | B\} = \frac{0,3 \cdot 0,3}{0,18} = 0,500.$$

Следовательно, вероятнее всего некондиционный прибор был изготовлен на заводе № 3.

Доказательство формулы (1.18) в случае полной системы событий, состоящей из произвольного числа k событий, в точности повторяет доказательство формулы (1.18). В таком общем виде формулу

$$P\{A_i | B\} = \frac{P\{A_i\}P\{B | A_i\}}{\sum_{j=1}^k P\{B | A_j\}P\{A_j\}} \quad (1.19)$$

принято называть *формулой Байеса*.

1.2. Непрерывное вероятностное пространство (аксиоматика А. Н. Колмогорова)

1.2.1. Специфика общего (непрерывного) случая вероятностного пространства

Ранее упоминалось о ситуациях, в которых множество всех возможных элементарных исходов (пространство элементарных событий Ω) может оказаться *более чем счетным*. Так, например, именно с *континуальным* пространством элементарных событий придется иметь дело, если

каждому элементарному исходу ω ; исследуемого случайного эксперимента (наблюдения) может быть поставлена в соответствие регистрация одной или нескольких числовых характеристик обследуемого объекта анализируемой совокупности, измеренных в физических единицах *непрерывной* природы (в единицах времени, длины, веса, температуры, давления и т. п.). Можно, правда, возразить, что, поскольку все измерения делаются с *ограниченной точностью*, реальное множество элементарных исходов все равно окажется не более чем счетным. Однако, с одной стороны, возможности точности измерений со временем совершенствуются и вместе с этим должна соответственно трансформироваться и структура рассматриваемого дискретного вероятностного пространства. С другой стороны, рассмотрение непрерывных моделей, отвечая физической сущности анализируемого явления, одновременно *расширяет аналитические возможности теории*, предоставляет исследователю более мощный математический аппарат: достаточно сопоставить возможности простого суммирования и интегрирования, аппарата разностных и дифференциальных уравнений и т. д.

Как же осуществляется переход от дискретного к непрерывному случаю в построении строгой математической теории вероятностей? Автоматический перенос всей схемы построения дискретного вероятностного пространства (см. п. 1.1) на непрерывный случай невозможен. Одно из принципиальных отличий непрерывного случая от дискретного заключается в том, что *в общем случае мы не можем объявить*, подобно тому как это делалось в дискретном вероятностном пространстве, *любое подмножество множества элементарных исходов Ω случайным событием*, т. е. событием, характеризующимся принципиальной возможностью его наблюдения в результате исследуемого случайного эксперимента. Другими словами, в общем вероятностном пространстве среди всех возможных подмножеств пространства элементарных событий Ω часть подмножеств характеризуется такой возможностью (и их принято называть случайными событиями или *измеримыми подмножествами Ω*), а другая часть — нет (подмножества Ω этого типа принято называть *неизмеримыми*).

Отмеченная особенность общего (непрерывного) случая, по-видимому, требует введения дополнительных определений и аксиом, относящихся к определению случайных событий и к правилам действий с ними и их вероятностями. Это и делается при аксиоматическом (теоретико-множественном) построении современной теории вероятностей, первое

строгое и полное изложение которой принадлежит А. Н. Колмогорову [Колмогоров]¹.

1.2.2. Случайные события, их вероятности и правила действий с ними (аксиоматический подход А. Н. Колмогорова)

Определим ту часть подмножеств пространства элементарных событий Ω , которая содержит *подмножества-события*. Схема определения случайного события A в общем случае подобна той, которая была принята в дискретном случае. Но если в той ситуации нам достаточно было определить в качестве исходных понятий элементарные события $\omega_1, \omega_2, \dots, \omega_N, \dots$ (и любое подмножество пространства элементарных событий объявлялось событием), то в общем случае мы в каждой конкретной реальной ситуации должны (из физических, содержательных соображений) определить дополнительно к Ω категорию подмножеств Ω , которые, очевидно, являются событиями. А затем любое случайное событие A определяется как некоторое производное от «очевидно событийных» подмножеств введенной категории.

Определение случайного события. Рассмотрим систему (конечную или счетную) подмножеств $A_1, A_2, \dots$ пространства элементарных событий Ω , каждое из которых является событием. Определим, что множество Ω , состоящее из всех элементарных событий, все дополнения $\bar{A}_i = \Omega - A_i$, пустое множество $\emptyset = \Omega - \Omega$ и всевозможные суммы $A = A_{i_1} + A_{i_2} + \dots + A_{i_n} + \dots$ также являются событиями. Отсюда непосредственно следует, что и произведения $\Pi = A_{i_1} A_{i_2} \dots A_{i_n} \dots$ также являются событиями, так как их дополнения $\bar{\Pi} = \bar{A}_{i_1} + \bar{A}_{i_2} + \dots$ в соответствии с данным определением являются событиями.

Будем обозначать в дальнейшем систему тех подмножеств пространства элементарных исходов Ω , которые являются событиями, буквой $\mathcal{C}$.

Аксиома. Каждому подмножеству-событию A из системы $\mathcal{C}$ соответствует неотрицательное и не превосходящее единицы число $p(A) = P\{A\}$, называемое вероятностью события A , причем задающая это соответствие числовая функция множеств $p(A)$ обладает следующими свойствами:

¹ Впервые интерпретация случайного события как множества с одновременной трактовкой его вероятности как *меры этого множества* была дана, по-видимому, в работе польского математика А. Ломницкого (Lomnicki A. Nouveau fondements du calcul des probabilités — Fundam. Math., 1923, 4). Однако теория вероятностей как *точная математическая теория* в надлежащем объеме впервые была построена А. Н. Колмогоровым.

а) $p(\Omega) = P\{\Omega\} = 1$;

б) если события $A_1, A_2, \dots, A_n, \dots$ несовместны, то

$$P\{A_1 + A_2 + \dots + A_n + \dots\} = P\{A_1\} + P\{A_2\} + \dots + P\{A_n\} + \dots$$

Из этой аксиомы непосредственно следует, в частности, связь между вероятностями прямого (A) и противоположного ($\bar{A}$) событий:

$$P\{\bar{A}\} = 1 - P\{A\}.$$

Аксиоматическое свойство б) вероятностей формулировалось и *доказывалось* в дискретном случае в виде *теоремы сложения вероятностей*.

Точно так же то, что называлось *теоремой умножения вероятностей* (и выводилось из определения и аксиомы в дискретном случае), в общем случае *принимается по определению*.

Определение условной вероятности. Условная вероятность $P\{A | B\}$ события A при условии, что уже имеет место событие B , определяется с помощью формулы

$$P\{AB\} = P\{A | B\}P\{B\}.$$

Правила действий с событиями и их вероятностями и, в частности, формула полной вероятности (1.14), формула Байеса (1.19), определение независимости для системы событий (1.12), (1.13) и другие, доказанные в дискретном случае, имеют место (и могут быть доказаны) и в случае общего вероятностного пространства.

Итак, для исчерпывающего описания механизма исследуемого случайного эксперимента в общем случае, т.е. для задания в этом случае вероятностного пространства, необходимо: 1) описать пространство элементарных событий Ω ; 2) описать систему C измеримых подмножеств этого пространства или таких подмножеств, которые должны быть *принципиально наблюдаемыми* (т.е. являются событиями); 3) каждому такому событию A из системы C поставить в соответствие неотрицательное число $P\{A\}$, называемое вероятностью события A , причем это соответствие должно удовлетворять требованиям а) и б) аксиомы (очевидно, такое соответствие P есть *числовая функция множества*, определенная на подмножествах системы C ; функции такого типа принято называть *вероятностными мерами*, определенными на системе подмножеств C). Поэтому если в дискретном случае для краткого символического описания вероятностного пространства достаточно было *пары* символов $\{\Omega, p\}$, то в общем случае для этой же цели требуется уже «тройка» $\{\Omega, C, P\}$.

Следует, однако, помнить, что всякая модель, всякая теория, и в том числе современная аксиоматическая концепция теории вероятностей, есть

лишь форма *приближенного* представления реальной действительности, форма, не свободная от недостатков. Чтобы предостеречь читателя от переоценки возможностей аксиоматической теоретико-вероятностной модели, рассмотрим несложный пример.

Практика долгосрочного социально-экономического и научно-технического прогнозирования широко использует различные формы *экспертных опросов*. Одной из таких форм является подход, при котором каждого из опрашиваемых экспертов просят субъективно оценить вероятность осуществления интересующего нас события в будущем¹. Подходя к моделированию этого процесса с позиций субъективистской школы вероятностей и соответственно интерпретируя каждого из опрашиваемых экспертов в качестве своеобразного «измерительного прибора», мы можем определить понятие случайного эксперимента как результат ответа эксперта на поставленный ему вопрос. В этом случае пространство элементарных исходов Ω , очевидно, должно состоять из всех точек отрезка $[0, 1]$. При конструировании системы S «наблюдаемых» подмножеств пространства Ω естественно было бы априори потребовать, чтобы любой отрезок $\Delta = [c_1, c_2]$, лежащий внутри отрезка $[0, 1]$ (т. е. $0 \leq c_1 < c_2 \leq 1$), принадлежал бы к категории *событий* (т. е. для любого отрезка $\Delta = [c_1, c_2]$ должна быть определена вероятность $P\{\Delta\} = P\{[c_1, c_2]\}$ того, что численный ответ наугад выбранного эксперта будет принадлежать этому отрезку). Но тогда в соответствии с определением случайного события в общем случае событиями будут не только отрезки, но и все, что можно получить из них применением к ним (взятым в счетном числе) суммирования и перемножения, а также взятием дополнения (т. е. противоположного события). Поэтому, выбирая произвольную точку c на отрезке $[0, 1]$ и рассматривая последовательность отрезков Δ_n вида $\Delta_n = [c - 1/n, c + 1/n]$, мы обнаруживаем, что *точка* должна быть событием, так как она является, как легко видеть, счетным произведением отрезков Δ_n . Итак, любая точка отрезка $[0, 1]$ — событие. Множество рациональных точек, как известно, складывается из счетного числа точек. Следовательно, это множество — тоже событие. Но множество иррациональных точек есть дополнение к множеству рациональных точек. Значит, и множество иррациональных точек — событие. Но вряд ли естественно, с *физической точки зрения*, считать наблюдаемыми (и, следовательно, физически различимыми) событиями факты принадлежности точки к множеству рациональных и к множеству иррациональных чисел.

¹ Подобная форма экспертиз используется, например, при построении сценариев будущего социально-экономического и промышленного развития стран.

Как видно из этого примера, *и использование общепринятой сейчас аксиоматической концепции теории вероятностей может приводить к плохо физически интерпретируемым выводам.* В данном примере мы не провели до конца построение вероятностного пространства, так как не определили (аксиоматически) способ вычисления вероятностей на отрезках, т.е. величин $P\{\Delta\} = P\{[c_1, c_2]\}$. Физически естественное аксиоматическое задание этих вероятностей также обусловлено спецификой реального комплекса условий, характеризующих наш случайный эксперимент.

Так, если представить, что мы находимся в самой «неблагоприятной» для прогноза ситуации (интересующее нас событие настолько удалено во времени и неопределенно или опрашиваемые эксперты настолько некомпетентны, что ответы экспертов оказываются приблизительно равномерно «разбросанными» по всей длине отрезка $[0, 1]$), то естественно предположить, что вероятности $P\{\Delta\}$ будут зависеть только от длины отрезка Δ и не будут зависеть от того, в каком именно месте отрезок располагается, и определить их соответственно с помощью соотношений¹

$$P\{[c_1, c_2]\} = c_2 - c_1. \quad (1.20)$$

Легко проверить, что заданные этим соотношением вероятности удовлетворяют свойствам а) и б) аксиомы.

Подчеркнем, кстати, на этом примере одно из существенных отличий широкого класса непрерывных вероятностных пространств от дискретных: вероятность осуществления любого элементарного события (т.е. любого *возможного* исхода) ω в данном примере равна нулю; однако для сколь угодно малого отрезка Δ вероятность $P\{\Delta\}$ всегда будет положительной (это непосредственно следует из (1.20)). Таким образом, в этом примере мы впервые встретились с казалось бы парадоксальной ситуацией, когда *события ω хотя и являются возможными, но обладают нулевой вероятностью.* Соответственно события $\bar{\omega} = \Omega - \omega$, являющиеся дополнением к событиям нулевой вероятности ω , хотя и не могут быть названы достоверными, но имеют вероятность осуществления, равную единице. О событиях типа $\bar{\omega}$ часто говорят как о событиях, происходящих «почти всегда». При более глубоком рассмотрении можно понять, что подобные ситуации в непрерывном вероятностном пространстве на самом деле не являются парадоксальными. Для пояснения этой мысли можно привести аналогию с физическим телом, имеющим определенную массу,

¹ О трех возможных подходах (априорном, апостериорно-частотном и апостериорно-модельном) к выработке и обоснованию подобных предположений о природе аксиоматически определяемых вероятностей см. п. 1.1.3.

в то время как ни одна из точек, составляющих это тело, сама массой не обладает. Очевидно, тело в этой аналогии играет роль события, точка — роль элементарного исхода, а масса — роль вероятности.

ВЫВОДЫ

1. Основаниями теоретико-вероятностного математического аппарата являются: *понятия* случайного эксперимента, его возможного исхода и пространства элементарных событий: *аксиома* о существовании и нормировке вероятностей элементарных событий; *определение* случайного события и способа вычисления его вероятности.

2. Способ построения современной строгой вероятностной теории *аксиоматический*, причем для построения дискретного вероятностного пространства, т. е. для модельного математического описания механизма случайного эксперимента, имеющего лишь конечное или счетное множество возможных элементарных исходов, достаточно постулировать одну аксиому (о существовании и нормировке вероятностей элементарных исходов) и одно определение (о способе вычисления вероятности любого события).

3. Термины «*механизм случайного эксперимента*», «*реальный комплекс условий, индуцирующий исследуемый статистический ансамбль*» и «*вероятностное пространство*» являются синонимами и могут быть математически заданы в дискретном случае с помощью описания всех возможных элементарных исходов и сопоставления с каждым из них вероятности своего появления (с помощью аналитического задания, таблично, графически, алгоритмически).

4. *Главная сложность* построения вероятностного пространства, соответствующего исследуемому реальному комплексу условий, — в конкретном задании вероятностей элементарных событий в дискретном случае или вероятностной меры — в непрерывном. Из трех возможных подходов к решению этой задачи — *априорного*, *апостериорно-частотного* и *апостериорно-модельного* — последний является наиболее легко практически реализуемым и наиболее эффективным.

5. Основные правила действий в дискретном вероятностном пространстве задаются *теоремами сложения и умножения вероятностей, формулами полной вероятности и Байеса*.

6. В общем (непрерывном) вероятностном пространстве в отличие от дискретного среди подмножеств пространства элементарных событий Ω могут быть такие, для которых не существует принципиальной возможности их наблюдения в результате исследуемого случайного эксперимента («ненаблюдаемые» или «неизмеримые» подмножества). Такие подмноже-

ства не могут быть названы событиями, так как если A — событие, то мы должны иметь возможность сказать, наступило оно или не наступило в результате эксперимента (в этом смысле оно «наблюдаемо»); только тогда можно говорить об относительной частоте его наступления в серии экспериментов, а следовательно, и о вероятности $P\{A\}$.

7. Отмеченная в предыдущем пункте особенность общего вероятностного пространства требует введения дополнительных определений и аксиом, относящихся к определению случайных событий и к правилам действий с их вероятностями. Современная аксиоматическая концепция теории вероятностей (впервые полно и строго изложенная А. Н. Колмогоровым в 1933 г.) строит общее вероятностное пространство, отправляясь от определения случайного события (с помощью перечисления допустимых теоретико-множественных комбинаций над подмножествами, априори являющимися событиями) и аксиомы о вероятностях как о неотрицательных и ограниченных единицей числовых функциях, аргументами которых являются подмножества-события. Эта концепция не противоречит рассмотренному ранее способу построения дискретного вероятностного пространства (она включает в себя этот способ в качестве частного случая и соответственно сохраняет все правила действий с вероятностями и событиями) и обуславливает возможность физической интерпретации вероятности события как относительной частоты его появления в достаточно длинной серии экспериментов.

8. Использование аксиоматической концепции теории вероятностей может в некоторых случаях, как и всякая другая модель, приводить к плохо физически интерпретируемым выводам.

9. В общем (непрерывном) вероятностном пространстве в отличие от дискретного *могут существовать возможные события, обладающие нулевой вероятностью появления*. Соответственно противоположные к ним события (их дополнения) хотя и не могут быть названы достоверными, но имеют вероятность осуществления, равную единице.

ГЛАВА 2. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

2.1. Определение и примеры случайных величин

Рассматривая приведенные выше примеры случайных экспериментов (см. п. 1.1.2, примеры 1.1–1.7), мы видим, что в большинстве случаев результат случайного эксперимента может быть описан одним или несколькими числами. Так, в примерах 1.2, 1.5 и 1.7 эти результаты означают соответственно число очков, выпавших при бросании игральной кости; число дефектных изделий, обнаруженных при качественном контроле N случайно отобранных из массовой продукции изделий; число сбоев автоматической линии за наугад выбранную рабочую смену. В примере 1.3 (четыrehкратное бросание игральной кости) результат каждого случайного эксперимента может быть записан четверкой чисел, или четырехмерным вектором, а в примере 1.6 (проверка основной и дополнительной выборок изделий объема соответственно N_1 и N_2 , случайно отобранных из продукции массового производства) — парой чисел, или *двумерным вектором*. Даже в примерах 1.1 и 1.4, на первый взгляд не связанных с регистрацией числовых характеристик, можно для удобства закодировать соответствующие случайные исходы, приписывая, например, исходам «аверс» (пример 1.1) и «изделие годно» (пример 1.4) числовую метку «ноль», а исходам «реверс» и «изделие дефектно» — *числовую метку «единица»*. Продолжая наши примеры и рассматривая в рамках теоретико-вероятностной схемы регистрацию одного или одновременно нескольких интересующих нас свойств (выраженных реальными или условно закодированными числами) у каждого из анализируемых объектов, мы приходим к общей схеме, в которой понятие случайного эксперимента реализуется в регистрации на каждом из таких случайно отобранных объектов набора числовых характеристик

$$\xi = (\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(p)}), \quad p \geq 1. \quad (2.1)$$

Какие именно числовые значения ξ мы будем иметь в результате данного конкретного эксперимента, зависит от множества не поддающихся учету случайных факторов и однозначно определяется в конечном счете осуществившимся в результате данного случайного эксперимента элементарным исходом ω (а это означает, что ξ является числовой (в случае $p = 1$) или векторной (в случае $p \geq 2$) функцией аргумента ω). Таким образом, мы приходим к следующему определению случайной величины (иногда пользуются равнозначным термином «случайный признак» или просто «признак»).

Таблица 2.1

Анализируемые свойства (характеристики) объектов $\xi^{(j)}$	Анализируемые объекты (семьи)			
	1	2	...	n
Социально-демографические и экономические характеристики семьи				
$\xi^{(1)}$ — социальная принадлежность	$x_1^{(1)}$	$x_2^{(1)}$	...	$x_n^{(1)}$
$\xi^{(2)}$ — профессия главы семьи	$x_1^{(2)}$	$x_2^{(2)}$	...	$x_n^{(2)}$
$\xi^{(3)}$ — качество жилищных условий	$x_1^{(3)}$	$x_2^{(3)}$	...	$x_n^{(3)}$
$\xi^{(4)}$ — размер (число членов) семьи	$x_1^{(4)}$	$x_2^{(4)}$	...	$x_n^{(4)}$
$\xi^{(5)}$ — количество детей	$x_1^{(5)}$	$x_2^{(5)}$	...	$x_n^{(5)}$
$\xi^{(6)}$ — среднедушевой доход	$x_1^{(6)}$	$x_2^{(6)}$	...	$x_n^{(6)}$
Характеристики «поведения» семьи (структура потребления)				
$\xi^{(7)}$ — расходы на питание	$x_1^{(7)}$	$x_2^{(7)}$	...	$x_n^{(7)}$
$\xi^{(8)}$ — расходы на промышленные товары текущего пользования	$x_1^{(8)}$	$x_2^{(8)}$	...	$x_n^{(8)}$
$\xi^{(9)}$ — расходы на предметы роскоши и длительного пользования	$x_1^{(9)}$	$x_2^{(9)}$	...	$x_n^{(9)}$
$\xi^{(10)}$ — расходы на услуги	$x_1^{(10)}$	$x_2^{(10)}$	...	$x_n^{(10)}$
$\xi^{(11)}$ — прочие расходы, включая сбережения	$x_1^{(11)}$	$x_2^{(11)}$	...	$x_n^{(11)}$

Случайной величиной называется поддающаяся измерению скалярная ($p = 1$) или векторная ($p \geq 2$) величина определенного физического смысла, значения (компоненты) которой подвержены некоторому не-

контролируемому разбросу при повторениях исследуемого эксперимента (наблюдения, процесса). Можно сказать, что случайная величина ξ — это функция, определенная на множестве элементарных событий, т. е. $\xi = \xi(\omega)$.

Пример 2.1. В табл. 2.1 приведен еще один пример *векторной* (или *многомерной*) случайной величины вместе с соответствующей ей общей формой регистрации серии наблюдений (результатов случайных экспериментов).

Обозначения случайной величины (которая по существу определяет лишь *перечень анализируемых характеристик* и по сложившейся в специальной литературе традиции чаще всего обозначается с помощью одной из букв греческого алфавита — ξ, η, ζ, ν и т. д.) отличаются от обозначений ее *наблюденных значений*.

В табл. 2.1 и в дальнейшем эти *наблюденные значения* в целях общности обозначаются с помощью строчных букв латинского алфавита (в таблице — с помощью буквы x) с верхним индексом, указывающим номер зафиксированной характеристики, и с нижним индексом, определяющим номер эксперимента или объекта, в котором эта характеристика зарегистрирована; но в любом случае нужно помнить, что за этими символами «скрываются» реальные *числовые значения* соответствующих характеристик или их *условные числовые метки*. Так, очевидно, в табл. 2.1 первые три строки будут состоять из условных числовых меток, а последующие восемь — из числовых значений, измеренных в определенной шкале и имеющих четкий физический смысл.

2.2. Возможные и наблюдаемые значения случайной величины

Поскольку в соответствии с одним из определений случайная величина — это функция, определенная на множестве элементарных исходов, то ее *возможные значения* и их общее число определяются структурой соответствующего пространства Ω элементарных событий: каждому элементарному событию ω соответствует свое возможное значение ξ . При этом, правда, может быть, что *нескольким элементарным исходам соответствует одно и то же возможное значение анализируемой случайной величины*, так что, вообще говоря, в конечном дискретном вероятностном пространстве число возможных значений случайной величины всегда меньше или равно числу различных элементарных исходов. Так, в приведенных выше примерах из п. 1.1.2 мы имеем: в примерах 1.1 и 1.4 всего по два возможных «значения», соответствующих элементарным ис-

ходам, — «аверс» и «реверс» (в примере 1.1) и «изделие годно» и «изделие дефектно» (в примере 1.4); в примере 1.2 — шесть первых положительных натуральных чисел задают одновременно и все элементарные исходы, и все возможные значения анализируемой случайной величины; в примере 1.7 возможными значениями числа сбоев автоматической линии за смену являются все неотрицательные целые числа. Что касается примеров 1.3, 1.5 и 1.6, то в них мы как раз оказываемся в ситуации, при которой число возможных значений анализируемой случайной величины будет меньше общего числа элементарных событий. Поясним это на примере 1.3 (примеры 1.5 и 1.6 могут быть проанализированы по аналогичной схеме). Пусть в примере 1.3 нас интересует число появившихся «шестерок» в серии из четырех бросаний игральной кости (независимо от порядка, в котором они будут появляться). Элементарный исход в этом примере может быть описан последовательностью из четырех символов, в которой на i -м месте записывают 1, если в результате i -го бросания игральной кости выпала «шестерка», и 0 — в противном случае. Тогда с точки зрения классификации элементарных событий последовательности (0, 1, 0, 0) и (0, 0, 0, 1) дают разные элементарные исходы, в то время как с точки зрения анализируемой случайной величины оба эти элементарных события дают неразличимый результат: и в том, и в другом случае наблюдаемое значение анализируемой случайной величины равно единице («шестерка» появилась ровно один раз). Конечно, нам не удастся «пересчитать» все возможные значения случайной величины, определенной на множестве элементарных исходов *непрерывного* пространства элементарных событий Ω : их общее число образует континуум. Именно такого рода величины представлены компонентами $\xi^{(6)} \div \xi^{(11)}$ многомерной (векторной) случайной величины ξ , рассмотренной в примере 2.1.

Следует отличать *теоретически возможные значения* случайной величины (обозначим их $x_1^0, x_2^0, \dots, x_i^0, \dots$ в дискретном случае и просто x — в непрерывном) от *практически осуществившихся в экспериментах*, т. е. от *наблюденных ее значений* (последние обозначим $x_1, x_2, \dots, x_n$)¹.

2.3. Типы случайных величин

Общая классификация возможных типов случайных величин может быть представлена с помощью схемы рис. 2.1.

¹ Если интересующая нас случайная величина многомерная (или векторная, см. (2.1)), то обозначения сохраняются с заменой строчных латинских букв x_i^0, x, x_i соответствующими прописными X_i^0, X, X_i .

Если мы в качестве результата эксперимента (наблюдения) регистрируем одно число (примеры 1.1, 1.2, 1.4, 1.5 и 1.7 из п.1.1.2; см. также случай $p = 1$ в записи (2.1)), то соответствующую случайную величину принято называть *одномерной* или *скалярной*. Если же результатом каждого эксперимента (наблюдения) является регистрация *целого набора* интересующих нас характеристик (примеры 1.3, 1.6 и 2.1, а также случай $p \geq 2$ в общей записи (2.1)), то соответствующую случайную величину называют *многомерной* или *векторной*.



Рис. 2.1. Общая схема классификации основных типов случайных величин

Одномерную случайную величину называют *дискретной* или *непрерывной* в зависимости от того, в каком пространстве элементарных событий она определена — в дискретном или в непрерывном. Очевидно, во всех рассмотренных выше примерах 1.1 – 1.7, так же как и в первых пяти компонентах из примера 2.1 (табл. 2.1), мы имеем дело с дискретными случайными величинами. Как уже сказано выше, некоторые исследователи, отправляясь от *ограниченности наших практических возможностей точности измерений*, предлагают вообще обходиться только дискретными вероятностными пространствами и соответственно только дискретными случайными величинами. Действительно, даже при измерении непре-

рывных по своей природе величин (длины, веса, температуры, давления и т. д.) всегда существует обусловленная разрешающей способностью используемого «измерительного прибора» максимально различимая единица измерения, своеобразный неразложимый квант, в целом числе которых и будет представлено в конечном счете наше измерение. Однако аналитические возможности непрерывных математических моделей, практика их непосредственного использования говорят за то, что они являются эффективным прикладным аппаратом применительно не только к непрерывным по своей физической природе случайным величинам, но и к таким дискретным, множество возможных значений которых достаточно велико (несколько десятков и более).

В зависимости от своей природы, своего назначения одномерные дискретные случайные величины подразделяются на количественные, ординальные (или порядковые) и номинальные (или классификационные).

Количественная случайная величина позволяет измерять *степень проявления* анализируемого свойства обследуемого объекта в определенной шкале (см. примеры 1.2, 1.3, 1.5, 1.6, 1.7, а также компоненты $\xi^{(4)} \div \xi^{(11)}$ в примере 2.1).

Ординальная (порядковая) случайная величина позволяет *упорядочивать обследуемые в ходе случайных экспериментов (наблюдений) объекты* по степени проявления в них анализируемого свойства. Исследователь обращается к ординальным случайным величинам в ситуациях, когда шкала, в которой можно было бы количественно измерить степень проявления анализируемого свойства, объективно не существует или ему не известна. В табл. 2.1 случайная величина $\xi^{(3)}$ — качество жилищных условий — предусматривает четыре возможные градации (категории качества): «плохое», «удовлетворительное», «хорошее» и «очень хорошее». Приписав каждой из обследованных семей (в соответствии с принятыми нормативными правилами) одну из градаций, мы тем самым получаем возможность упорядочить обследованные семьи по этому свойству. Общее число градаций ординального признака может быть меньше, равно и даже больше числа обследованных объектов (случайных экспериментов).

Номинальная (классификационная) случайная величина позволяет *разбивать обследуемые в ходе случайных экспериментов (наблюдений) объекты на не поддающиеся упорядочению* однородные по анализируемому свойству классы. Если исследователю наряду с анализируемым свойством известны все возможные его градации (не поддающиеся упорядочению), вместе с правилом отнесения обследованного в ходе случайного эксперимента (наблюдения) объекта к одной из этих градаций, то соответствующую номинальную величину принято называть *категоризованной*.

Именно к таким признакам относятся случайные величины $\xi^{(1)}$ — социальная принадлежность семьи и $\xi^{(2)}$ — профессия главы семьи из табл. 2.1. Если условия эксперимента таковы, что его элементарным исходом является так называемое парное сравнение δ_{ij} , задающее меру сходства (или различия) по анализируемому свойству объектов с номерами i и j из обследуемой совокупности, то такую номинальную случайную величину будем называть *некатегоризованной*, а ее наблюдаемые значения представляют соответственно так называемой *матрицей смежности*¹:

$$\Delta = (\delta_{ij}), \quad i, j = \overline{1, n}.$$

Примером некатегоризованного номинального признака может служить случайная величина, индуцированная случайным экспериментом на различных парах семей, результатом которого является отнесение или неотнесение каждой пары к общему классу с точки зрения однородности (сходства) их потребительского поведения.

2.4. Одномерные и многомерные (совместные) законы распределения вероятностей случайных величин

Мы уже знаем (см. п. 1.1.3), что для полного описания вероятностного пространства или, что то же, для *исчерпывающего задания интересующей нас случайной величины* недостаточно задать лишь пространство элементарных событий Ω (и тем самым описать множество теоретически возможных значений анализируемой случайной величины). К этому необходимо добавить также: в *дискретном случае* — правило сопоставления с каждым возможным значением X_i^0 случайной величины ξ вероятности его появления $p_i = P\{\xi = X_i^0\}$; в *непрерывном случае* — правило сопоставления с каждой измеримой² областью ΔX возможных значений случайной величины ξ вероятности $p(\Delta X) = P\{\xi \in \Delta X\}$ события, заключающегося в том, что в случайном эксперименте реализуется одно из

¹ В наиболее распространенном частном случае элементы δ_{ij} могут принимать лишь два значения: 1, если объекты i и j отнесены в результате случайного эксперимента (наблюдения) к одной градации (или к одному классу), и 0 — в противном случае.

² Область возможных значений ΔX называется *измеримой*, если элементарные исходы ω , соответствующие значениям, вошедшим в эту область, образуют *измеримое подмножество* или *событие*, т. е. подмножество, принадлежащее системе $\mathcal{C}$ всех возможных событий (см. п. 1.2.1 и 1.2.2).

возможных значений, принадлежащих заданной области ΔX . Это правило, позволяющее устанавливать соответствия вида

$$\begin{aligned} X_i^0 &\rightarrow p_i = P\{\xi = X_i^0\}; \\ \Delta X &\rightarrow p(\Delta X) = P\{\xi \in \Delta X\}, \end{aligned} \quad (2.2)$$

принято называть *законом распределения вероятностей* исследуемой случайной величины ξ . Прозрачное пояснение такой терминологии мы получаем в рамках дискретного вероятностного пространства, поскольку в этом случае речь идет о *правиле распределения суммарной единичной вероятности* (т. е. вероятности достоверного события) между отдельными возможными значениями X_i^0 ($i = 1, 2, \dots$).

Очевидно, задание закона распределения вероятностей, т. е. соответствий типа (2.2), может осуществляться с помощью таблиц (только в дискретном случае), графиков, а также с помощью функций и алгоритмически (об основных формах задания законов распределения и примерах их модельной, т. е. аналитической, записи см. гл. 3).

Приведем примеры табличного и графического задания законов распределения вероятностей.

Тщательный статистический анализ засоренности партий дефектными изделиями (пример 1.5) позволил построить следующее распределение вероятностей для случайной величины ξ , выражающей число дефектных изделий, обнаруженных при контроле партии, состоящей из $N = 30$ изделий, случайно отобранных из продукции массового производства (табл. 2.2):

Таблица 2.2

i	1	2	3	4	5	6	7	8	9	10	11	12	...	31
x_i^0	0	1	2	3	4	5	6	7	8	9	10	11	...	30
$1000 \cdot p_i$	42	141	228	236	177	102	47	18	6	1	0	0	...	0

Значения вероятностей, приведенные в табл. 2.2, даны с точностью до третьего десятичного знака, поэтому то, что суммирование представленных в таблице вероятностей дает 0,998 (вместо единицы), легко объяснить: недостающие 0,002 как-то «размазаны» между возможными значениями 10, 11, ..., 30, но на каждое отдельное возможное значение приходится вероятность, меньшая 0,0005.

Тот же закон распределения может быть представлен графически (рис. 2.2).

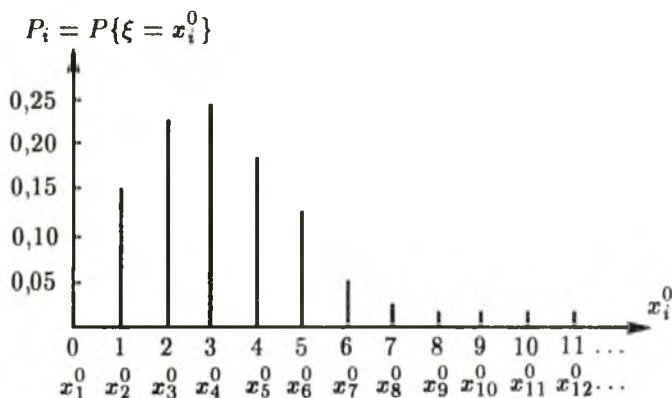


Рис. 2.2. Графическое задание закона распределения вероятностей для числа дефектных изделий, обнаруженных в наугад извлеченной партии, состоящей из 30 изделий массового производства

Геометрическое изображение закона распределения вероятностей дискретной случайной величины часто называют *полигоном распределения* или *полигоном вероятностей*.

В качестве другого примера рассмотрим фрагмент табл. 2.1, выбрав из одиннадцати представленных в ней компонент только две: качество жилищных условий $\xi^{(3)}$ и среднедушевой доход $\xi^{(6)}$. Еще более упростим рассматриваемую схему, перейдя от по существу *непрерывной* случайной величины $\xi^{(6)}$ к ее дискретному аналогу $\tilde{\xi}^{(6)}$, т. е. отказываясь от точного знания среднедушевого дохода каждой семьи и ограничиваясь лишь тремя возможными градациями: семья имеет низкий доход (градация $\tilde{x}_1^{(6)}$), средний доход (градация $\tilde{x}_2^{(6)}$) и высокий доход (градация $\tilde{x}_3^{(6)}$). С учетом четырех градаций качества жилищных условий: $x_1^{(3)}$ — качество низкое; $x_2^{(3)}$ — качество удовлетворительное; $x_3^{(3)}$ — качество хорошее и $x_4^{(3)}$ — качество очень хорошее — и проведенного вероятностно-статистического анализа получаем следующий закон распределения вероятностей *двумерной* случайной величины $(\tilde{\xi}^{(6)}, \xi^{(3)})$:

Соответствующий двумерный полигон распределения представлен на рис. 2.3.

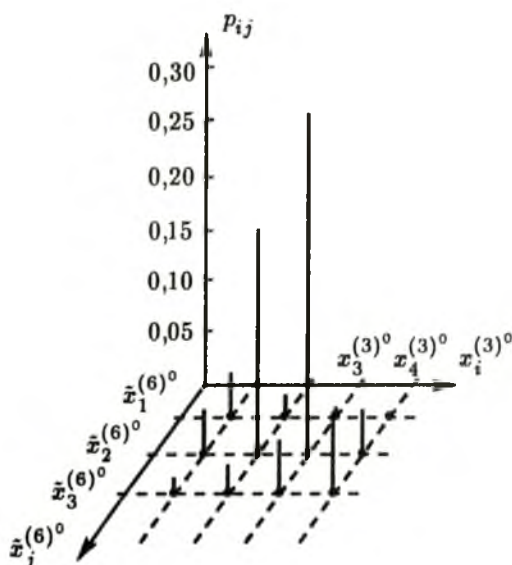


Рис. 2.3. Полигон двумерного распределения семей по качеству жилищных условий ($\xi^{(3)}$) и уровню дохода ($\tilde{\xi}^{(6)}$): $p_{ij} = P\{\tilde{\xi}^{(6)} = \tilde{x}_i^{(6)0}, \xi^{(3)} = x_j^{(3)0}\}$

Закон распределения вероятностей многомерной случайной величины называют многомерным или совместным. Если каждая из компонент $\xi^{(k)}$ ($k = 1, 2, \dots, p$; см. (2.1)) анализируемого многомерного признака ξ дискретна и имеет конечное число m_k всех возможных значений, то, очевидно, общее число возможных «значений» случайного вектора ξ будет $m = m_1 \cdot m_2 \dots m_p$. В этом случае вместо *общей индексации* всех возможных многомерных значений X_l^0 ($l = 1, 2, \dots, m$) удобнее пользоваться *p-мерной индексацией* вида $ij \dots q$, где первый индекс i определяет номер возможного значения по первой компоненте, второй индекс j — по второй компоненте и т. д. Тогда $X_{ij \dots q}^0$ будет означать возможное значение ξ , полученное сочетанием i -го возможного значения компоненты $\xi^{(1)}$, j -го возможного значения компоненты $\xi^{(2)}$ и т. д. до q -го возможного значения компоненты $\xi^{(p)}$, а вероятности $P\{\xi = X_{ij \dots q}^0\}$ удобно обозначать $p_{ij \dots q}$. Таким образом, в табл. 2.3 представлены вероятности $p_{ij} = P\{\tilde{\xi}^{(6)} = \tilde{x}_i^{(6)0}, \xi^{(3)} = x_j^{(3)0}\}$.

Таблица 2.3

i	$j:$	1	2	3	4	$p_i.$
	$x_j^{(3)0}$					
	$\bar{x}_i^{(6)0}$	$x_1^{(3)0}$	$x_2^{(3)0}$	$x_3^{(3)0}$	$x_4^{(3)0}$	
1	$\bar{x}_1^{(6)0}$	0,06	0,03	0,01	0,00	0,10
2	$\bar{x}_2^{(6)0}$	0,05	0,25	0,35	0,05	0,70
3	$\bar{x}_3^{(6)0}$	0,01	0,02	0,07	0,10	0,20
$p_{.j}$		0,12	0,30	0,43	0,15	1,00

При анализе многомерных (совместных) распределений часто бывает достаточно знать закон распределения *лишь для какой-то части компонент* анализируемого векторного признака. Так, многомерная случайная величина ξ , рассмотренная в табл. 2.1, естественно разбивается на два подвектора: $\xi_1 = (\xi^{(1)}, \dots, \xi^{(6)})$, описывающий социально-демографические и экономическую характеристики семьи, и $\xi_2 = (\xi^{(7)}, \dots, \xi^{(11)})$, описывающий структуру семейного потребления. Поэтому, если нас интересует лишь социально-демографо-экономическая структура исследуемой совокупности семей, то предметом нашего анализа будет закон распределения вероятностей только по компонентам $\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(6)}$ подвектора ξ_1 .

Частный (маржинальный) закон распределения подвектора ξ_1 анализируемой многомерной случайной величины $\xi = (\xi_1, \xi_2)$ описывает распределение вероятностей признака ξ_1 в ситуации, когда на значения другой части компонент ξ_2 не накладывается никаких условий. В дискретном случае соответствующие вероятности определяются по формулам:

$$p_{i.} = P\{\xi_1 = X_i^{(1)0}\} = \sum_j P\{\xi_1 = X_i^{(1)0}, \xi_2 = X_j^{(2)0}\}; \quad (2.3)$$

$$p_{.j} = P\{\xi_2 = X_j^{(2)0}\} = \sum_i P\{\xi_1 = X_i^{(1)0}, \xi_2 = X_j^{(2)0}\}, \quad (2.3')$$

где $X_i^{(1)0}$ и $X_j^{(2)0}$ — i -е и j -е возможные значения векторных признаков соответственно ξ_1 и ξ_2 , а суммирование производится по всем возможным значениям $X_j^{(2)0}$ в формуле (2.3) и по всем возможным значениям $X_i^{(1)0}$ в формуле (2.3').

Формулы (2.3) и (2.3') получаются как непосредственные следствия

теоремы сложения вероятностей (1.9''), если принять во внимание следующие очевидные связи между интересующими нас событиями:

$$\begin{aligned}\{\xi_1 = X_i^{(1)0}\} &= \{\xi_1 = X_i^{(1)0}; \xi_2 = X_1^{(2)0}\} \\ &\quad + \{\xi_1 = X_i^{(1)0}; \xi_2 = X_2^{(2)0}\} \\ &\quad + \{\xi_1 = X_i^{(1)0}; \xi_2 = X_3^{(2)0}\} + \dots \\ \{\xi_2 = X_j^{(2)0}\} &= \{\xi_1 = X_1^{(1)0}; \xi_2 = X_j^{(2)0}\} \\ &\quad + \{\xi_1 = X_2^{(1)0}; \xi_2 = X_j^{(2)0}\} \\ &\quad + \{\xi_1 = X_3^{(1)0}; \xi_2 = X_j^{(2)0}\} + \dots\end{aligned}$$

В рассматриваемом примере (см. табл. 2.3) частные распределения $p_{i.} = P\{\tilde{\xi}^{(6)} = \tilde{x}_i^{(6)0}\}$ и $p_{.j} = P\{\xi^{(3)} = x_j^{(3)0}\}$ подсчитаны по формулам (2.3) и (2.3') и задают соответственно распределение семей отдельно по уровню дохода и по качеству жилищных условий (они приведены соответственно в последнем столбце и в последней строке табл. 2.3).

Условный закон распределения подвектора ξ_1 анализируемой многомерной случайной величины $\xi = (\xi_1, \xi_2)$ при условии, что значение другого подвектора ξ_2 зафиксировано на уровне $X_j^{(2)0}$, вычисляется по формуле

$$\begin{aligned}p_{i.}(X_j^{(2)0}) &= P\{\xi_1 = X_i^{(1)0} \mid \xi_2 = X_j^{(2)0}\} \\ &= \frac{P\{\xi_1 = X_i^{(1)0}; \xi_2 = X_j^{(2)0}\}}{P\{\xi_2 = X_j^{(2)0}\}} = \frac{p_{ij}}{p_{.j}}.\end{aligned}\quad (2.4)$$

Аналогично

$$\begin{aligned}p_{.j}(X_i^{(1)0}) &= P\{\xi_2 = X_j^{(2)0} \mid \xi_1 = X_i^{(1)0}\} \\ &= \frac{P\{\xi_1 = X_i^{(1)0}; \xi_2 = X_j^{(2)0}\}}{P\{\xi_1 = X_i^{(1)0}\}} = \frac{p_{ij}}{p_{i.}}.\end{aligned}\quad (2.4')$$

Формулы (2.4) и (2.4') получаются как простые следствия теоремы умножения вероятностей (1.11).

Так, например, если нас интересует условное распределение группы семей с высоким доходом по качеству жилищных условий, т. е. распределение $p_{.j}(\tilde{x}_3^{(6)0}) = P\{\xi^{(3)} = x_j^{(3)0} \mid \tilde{\xi}^{(6)} = \tilde{x}_3^{(6)0}\}$, то вычисления по (2.4) на

основе данных табл. 2.3 дают:

$$p_1(\tilde{x}_3^{(6)0}) = \frac{0,01}{0,20} = 0,05;$$

$$p_2(\tilde{x}_3^{(6)0}) = \frac{0,02}{0,20} = 0,10;$$

$$p_3(\tilde{x}_3^{(6)0}) = \frac{0,07}{0,20} = 0,35;$$

$$p_4(\tilde{x}_3^{(6)0}) = \frac{0,10}{0,20} = 0,50,$$

что означает, в частности, что из всей совокупности *высокодозодных* семей в плохих жилищных условиях проживает 5%, в удовлетворительных — 10%, в хороших — 35% и в очень хороших — 50% (то, что речь идет *только о высокодозодных семьях*, определяется условием $\{\tilde{\xi}^{(6)} = \tilde{x}_3^{(6)0}\}$).

2.5. Способы задания закона распределения: функция распределения и функция плотности вероятности

2.5.1. Функция распределения вероятностей одномерной случайной величины

Как установлено выше (см. п. 2.4), всякая генеральная совокупность (случайная величина) определяется своим законом распределения вероятностей $p(\Delta X)$. Поскольку интересующие нас области ΔX могут быть в общем случае подмножествами общей природы, то возникает вопрос: каковы те способы задания числовых функций p , определенных на подмножествах ΔX , которые были бы достаточно удобны в плане конструктивном, практическом?

Оказывается, для описания распределений *одномерных* случайных величин ξ достаточно задать способ вычисления вероятностей $p(\Delta X) = P\{\xi \in \Delta X\}$ *лишь для подмножеств ΔX некоторого специального вида*, а именно лишь для полузамкнутых слева интервалов вида

$$\Delta X = [x_{\min}, x),$$

где $x_{\min}$ — минимально возможное значение исследуемой случайной величины ξ (оно может быть равно и $-\infty$), а x — любое «текущее» (т. е. задаваемое нами) возможное значение ξ . Вероятность же $p([x_{\min}, x)) =$

$P\{\xi < x\}$ однозначно определяется заданием правого конца интервала, т. е. числа x , а потому может интерпретироваться как обычная функция от одного числового аргумента x .

Функцией распределения вероятностей (накопленной частотой) $F_\xi(x)$ случайной величины ξ называют функцию, ставящую в соответствие любому заданному значению x величину вероятности события $\{\xi < x\}$, т. е.

$$F_\xi(x) = P\{\xi < x\}. \quad (2.5)$$

В дальнейшем, если это не будет вызывать недоразумений, будем опускать нижний индекс ξ у функции F и называть ее просто «функцией распределения».

Рассмотрим поведение функции распределения. Во-первых, отметим, что в *дискретном* случае событие $A(x) = \{\xi < x\}$ состоит из всех элементарных событий $\omega_i = \{\xi = x_i^0\}$, таких, что $x_i^0 < x$. Поэтому в соответствии с определением вероятности составного события (см. п. 1.1.3) имеем

$$F(x) = P\{\xi < x\} = \sum_{i: x_i^0 < x} P\{\xi = x_i^0\} = \sum_{i: x_i^0 < x} p_i \quad (2.5')$$

(суммирование в правых частях (2.5') проводится по всем тем i , для которых $x_i^0 < x$).

Из (2.5') видно, что значения функции $F(x)$ изменяются при увеличении аргумента x *скачками*, а именно при «переползании» величины x через очередное возможное значение x_i^0 функция $F(x)$ скачком увеличивает свое значение на величину $p_i = P\{\xi = x_i^0\}$.

Несколько иную картину мы будем наблюдать, анализируя поведение функции распределения $F_\xi(x)$ в случае *непрерывного* исследуемого признака ξ . Большинство представляющих практический интерес непрерывных случайных величин обладает тем свойством, что для любого отрезка Δx вероятности $P\{\xi \in \Delta x\}$ стремятся к нулю по мере стремления к нулю длины этого отрезка, и, следовательно, *вероятности отдельных возможных значений x равны нулю* (конкретный пример такого рода приведен в п. 1.2.2 в задаче с экспертным оцениванием вероятности интересующего нас события). Нетрудно понять, что для таких случайных величин их функции распределения оказываются *непрерывными*. На рис. 2.4, а–г представлены графики функций распределения случайных величин, рассмотренных соответственно в примерах 1.1, 1.2, 1.5 (с учетом табл. 2.2) и в примере с экспертным оцениванием вероятности интересующего нас события (п. 1.2.2).

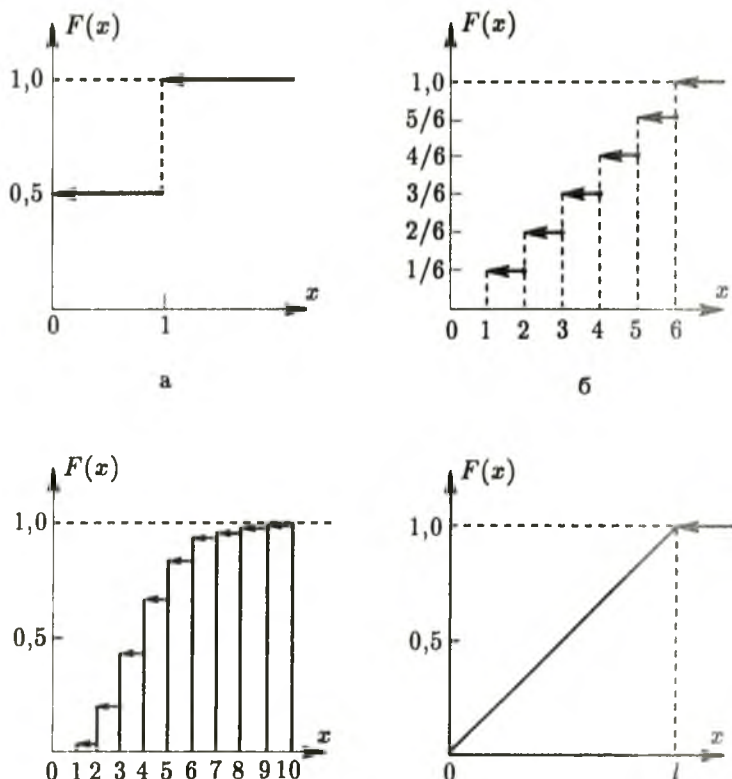


Рис. 2.4. Графики функций распределения для: а — оцифрованного результата подбрасывания монеты (ноль соответствует аверсу, единица — реверсу); б — числа очков, выпадающих при бросании правильной игральной кости; в — числа дефектных изделий, обнаруженных в наугад выбранной партии, состоящей из 30 изделий (см. табл. 2.2); г — экспертной оценки вероятности интересующего нас события (при полной некомпетентности экспертов), см. пример из п. 1.2.2. Стрелки на графиках означают, что точки, в которые они направлены, не включаются в состав обозначенных ими интервалов группирования

Из определения функции распределения непосредственно вытекают следующие ее основные свойства:

- $F_{\xi}(x)$ — неубывающая функция аргумента x ;
- $F_{\xi}(x) = 0$ для всех $x \leq x_{\min}$;
- $F_{\xi}(x) = 1$ для всех $x > x_{\max}$ ($x_{\min}$ и $x_{\max}$ — соответственно минимальное и максимальное возможные значения исследуемой случайной величины ξ);

г) $P\{a \leq \xi < b\} = F_\xi(b) - F_\xi(a)$ для любых заданных значений a и b (для доказательства последнего свойства следует воспользоваться теоремой сложения вероятностей (см. п. 1.1.3), а также тем обстоятельством, что события $A = \{\xi < a\}$, $B = \{\xi < b\}$ и $C = \{a \leq \xi < b\}$ связаны между собой соотношением $B = A + C$).

2.5.2. Функция плотности вероятности одномерной случайной величины

В классе таких *непрерывных* случайных величин, функции распределений которых всюду непрерывны и дифференцируемы (а, как уже отмечалось, этот класс охватывает большинство представляющих практический интерес непрерывных случайных величин), другой удобной формой задания генеральной совокупности (исследуемой случайной величины ξ) является функция *плотности вероятности* $f_\xi(x)$, определяемая как предел

$$\begin{aligned} f_\xi(x) &= \lim_{\Delta \rightarrow 0} \frac{P\{\xi \in [x + \Delta]\}}{\Delta} \\ &= \lim_{\Delta \rightarrow 0} \frac{F_\xi(x + \Delta) - F_\xi(x)}{\Delta}, \end{aligned} \quad (2.6)$$

или, что то же,

$$f_\xi(x) = F'_\xi(x), \quad (2.6')$$

т. е. $f_\xi(x)$ — это производная функции распределения $F_\xi(x)$ в точке x . Из эквивалентных соотношений (2.6) и (2.6'), определяющих *функцию плотности* $f_\xi(x)$, вытекают непосредственно следующие ее свойства:

- а) $f_\xi(x) \geq 0$, так как функция $F_\xi(x)$ неубывающая;
- б) $P\{\xi \in [x, x + \Delta]\} \approx f_\xi(x) \cdot \Delta$ для малых отрезков Δ (следует из сравнения первых двух членов тождества (2.6));

$$\text{в) } P\{\xi \in [x_{\min}, x]\} = F_\xi(x) = \int_{x_{\min}}^x f_\xi(u) du \text{ для любых } x;$$

$$\text{г) } P\{\xi \in [x_0, x_0 + \Delta]\} = F_\xi(x_0 + \Delta) - F_\xi(x_0) = \int_{x_0}^{x_0 + \Delta} f_\xi(u) du \text{ для любых } x_0 \text{ и } \Delta;$$

$$\text{д) } P\{\xi \in [x_{\min}, x_{\max}]\} = \int_{x_{\min}}^{x_{\max}} f_\xi(x) dx \equiv 1.$$

Прокомментируем некоторые из этих свойств функции плотности.

Свойство б) позволяет пояснить вероятностный смысл функции плотности. Так, предположив для определенности область возможных значений $[x_{\min}, x_{\max}]$ исследуемой случайной величины ξ конечной и разбив ее

на одинаковые и достаточно мелкие интервалы группирования Δ с центрами

$$x_1^0 = x_{\min} + \frac{\Delta}{2}, \quad x_2^0 = x_1^0 + \Delta, \quad x_3^0 = x_2^0 + \Delta$$

и т. д., мы можем поставить в соответствие каждому i -му интервалу вероятность осуществления события

$$\xi \in \left[x_i^0 - \frac{\Delta}{2}, x_i^0 + \frac{\Delta}{2} \right],$$

приближенно равную в соответствии со свойством б) величине $f_\xi(x_i^0) \cdot \Delta$. Таким образом, по своему смыслу значения функции $f_\xi(x)$ пропорциональны вероятности того, что исследуемая случайная величина примет значение в непосредственной близости от точки x . Этот факт, в частности, может служить основанием к тому, что дискретным аналогом функции плотности в случае дискретной случайной величины является полигон частот, т. е. последовательность точек с координатами (x_i^0, p_i) . Отсюда же следует, что наиболее вероятным (модальным) значением исследуемой непрерывной случайной величины является такое ее возможное значение x_{mod} , в котором функция плотности достигает своего максимума, т. е. $f_\xi(x_{\text{mod}}) = \max_x f_\xi(x)$.

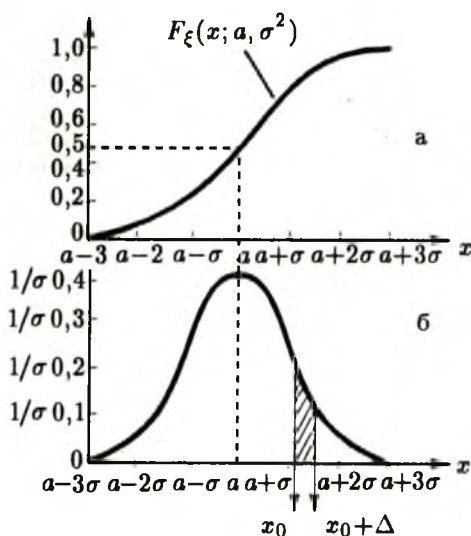


Рис. 2.5. Функции: (а) распределения $F_{\xi \text{ норм}}(x; a; \sigma^2)$ и (б) плотности $f_{\xi \text{ норм}}(x; a; \sigma^2)$ нормального закона

Геометрическая интерпретация свойства г) состоит в том, что вероятность события $\{\xi \in [x_0, x_0 + \Delta)\}$ оказывается (при любых заданных x_0 и Δ) равной площади «столбика» под кривой плотности $y = f_\xi(x)$ с основанием $[x_0, x_0 + \Delta)$.

На рис. 2.5 показаны функции распределения $F_{\xi_{\text{норм}}}(x; a; \sigma^2)$ и плотности $f_{\xi_{\text{норм}}}(x; a; \sigma^2)$ одного из распространенных законов распределения — *нормального* (подробнее см. п. 3.1 и 4.3). Заштрихованная площадь на рис. 2.5, б дает геометрически наглядное представление о величине вероятности $P\{\xi \in [x_0, x_0 + \Delta)\}$.

2.5.3. Многомерные функции распределения и плотности. Статистическая независимость случайных величин

Из вышеизложенного ясно, что вопрос об удобных способах задания закона распределения случайной величины особенно актуален в непрерывном случае: для описания «поведения» дискретной случайной величины ξ универсальной и одновременно конструктивной формой (при «не слишком большом» числе возможных значений исследуемой случайной величины) является *полигон* вероятностей, т. е. форма, при которой каждому возможному значению x_i^0 ставится в соответствие вероятность его осуществления $p_i = P\{\xi = X_i^0\}$. Поэтому сосредоточим теперь свое внимание на непрерывном случае. *Специфика многомерных схем в этом случае заключается в том, что в отличие от одномерного случая многомерная функция распределения*

$$F_{(\xi^{(1)}, \dots, \xi^{(p)})}(x^{(1)}, \dots, x^{(p)}) = P\{\xi^{(1)} < x^{(1)}, \dots, \xi^{(p)} < x^{(p)}\} \quad (2.7)$$

перестает быть исчерпывающей (по информативности) формой задания изучаемого закона распределения.

Многомерными аналогами конечных и полубесконечных отрезков (которые можно получить суммированием и пересечением полубесконечных отрезков вида $[-\infty, x)$) являются конечные и полубесконечные *гиперпараллелепипеды*. Именно для многомерных областей такого типа и определяет функция распределения (2.7) правило вычисления вероятностей. Однако, если в одномерном случае этого было достаточно для «работы» в соответствующем вероятностном пространстве, то в многомерном случае нас это уже не удовлетворяет. В частности, знание одной лишь формы (2.7) оказывается недостаточным для конструктивного решения такой важной для статистических приложений задачи, какой является задача описания закона распределения интересующих нас *преобразований* от исходных случайных величин $\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(p)}$ (общий подход к решению

этой проблемы описан в п. 4.4).

Поэтому для описания закона распределения многомерной величины $\xi = (\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(p)})$ в непрерывном случае используют функцию плотности вероятности $f_{\xi}(x^{(1)}, x^{(2)}, \dots, x^{(p)})$, которую определим как такую функцию $f(x^{(1)}, \dots, x^{(p)})$ от p переменных, что для любого (измеримого) подмножества A возможных значений ξ вероятность события $\{\xi \in A\}$ может быть вычислена с помощью соотношения

$$P\{\xi \in A\} = \int_A f_{\xi}(x^{(1)}, \dots, x^{(p)}) dx^{(1)} \dots dx^{(p)}, \quad (2.8)$$

где интегрирование ведется по данной области A в соответствующем p -мерном пространстве возможных значений ξ (т. е. знак интеграла в (2.8) определяет операцию p -кратного интегрирования).

Из данного определения функции плотности вероятности (фпв) непосредственно следует, что фпв и функция распределения (2.7) связаны между собой соотношениями¹:

$$\begin{aligned} F_{\xi}(x^{(1)}, \dots, x^{(p)}) \\ = \int_{-\infty}^{x^{(1)}} \int_{-\infty}^{x^{(2)}} \dots \int_{-\infty}^{x^{(p)}} f_{\xi}(u^{(1)}, u^{(2)}, \dots, u^{(p)}) du^{(1)} du^{(2)} \dots du^{(p)}, \end{aligned} \quad (2.7')$$

$$f_{\xi}(x^{(1)}, \dots, x^{(p)}) = \frac{\partial^p F(x^{(1)}, \dots, x^{(p)})}{\partial x^{(1)} \partial x^{(2)} \dots \partial x^{(p)}}. \quad (2.9)$$

Вероятностный смысл функции плотности тот же, что и в одномерном случае: вероятность осуществления значения случайной величины ξ , лежащего в некоторой малой окрестности $\Delta X = \{x^{(1)} \leq \xi^{(1)} < x^{(1)} + \Delta x^{(1)}, x^{(2)} \leq \xi^{(2)} < x^{(2)} + \Delta x^{(2)}, \dots, x^{(p)} \leq \xi^{(p)} < x^{(p)} + \Delta x^{(p)}\}$ точки $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})$, пропорциональна значению $f_{\xi}(X)$ функции плотности в этой точке и приблизительно равна, в частности, «элементу

¹ Так же, как и в одномерном случае, наше описание относится лишь к таким непрерывным случайным величинам, функция распределения (2.7) которых непрерывна по всем переменным внутри области возможных значений и имеет в этой области по всем своим переменным непрерывную частную производную. Как отмечалось, такие случайные величины составляют большинство в классе ситуаций, представляющих практический интерес.

вероятности» $f_{\xi}(X)\Delta X$, т. е.

$$\begin{aligned} P\{x^{(1)} \leq \xi^{(1)} < x^{(1)} + \Delta x^{(1)}, \dots, x^{(p)} \leq \xi^{(p)} < x^{(p)} + \Delta x^{(p)}\} \\ \approx f_{\xi}(x^{(1)}, \dots, x^{(p)}) \Delta x^{(1)} \dots \Delta x^{(p)}. \end{aligned} \quad (2.10)$$

Функции, определяемые соотношениями (2.7), или (2.7'), а также (2.9), называют соответственно *совместной* функцией распределения и *совместной* плотностью вероятности многомерного случайного признака $\xi = (\xi^{(1)}, \dots, \xi^{(p)})$.

Для описания *частного* закона распределения вероятностей некоторой *части* компонент $\xi_1 = (\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(s)})$, $s < p$, вектора ξ (см. п. 2.4.1) используются *частная (маргинальная) функция распределения* $F_{\xi_1}(x^{(1)}, \dots, x^{(s)})$ и *частная (маргинальная) плотность вероятности*, задаваемые соотношениями:

$$\begin{aligned} F_{\xi_1}(x^{(1)}, \dots, x^{(s)}) &= P\{\xi^{(1)} < x^{(1)}, \dots, \xi^{(s)} < x^{(s)}\} \\ &= P\{\xi^{(1)} < x^{(1)}, \dots, \xi^{(s)} < x^{(s)}, \xi^{(s+1)} < \infty, \dots, \xi^{(p)} < \infty\} \\ &= F_{\xi}(x^{(1)}, x^{(2)}, \dots, x^{(s)}, \infty, \dots, \infty); \end{aligned} \quad (2.11)$$

$$\begin{aligned} f_{\xi_1}(x^{(1)}, \dots, x^{(s)}) \\ = \int_{x^{(s+1)}} \int_{x^{(s+2)}} \dots \int_{x^{(p)}} f_{\xi}(x^{(1)}, \dots, x^{(s)}, x^{(s+1)}, \dots, x^{(p)}) \\ \times dx^{(s+1)} \dots dx^{(p)}, \end{aligned} \quad (2.12)$$

где под $\int_{x^{(q)}}$ понимается интегрирование по всему множеству возможных значений случайной величины $\xi^{(q)}$ (ср. с (2.3) и (2.3')).

Условная плотность вероятности $f_{\xi}(x^{(1)}, x^{(2)}, \dots, x^{(s)} \mid \xi_2 = C)$ случайного вектора $\xi_1 = (\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(s)})$ при условии, что значения другого подвектора $\xi_2 = (\xi^{(s+1)}, \xi^{(s+2)}, \dots, \xi^{(p)})$ зафиксированы на уровне $C = (c^{(s+1)}, c^{(s+2)}, \dots, c^{(p)})$ (т. е. при условии $\xi^{(s+1)} = c^{(s+1)}, \dots, \xi^{(p)} = c^{(p)}$), определяется аналогично условным вероятностям p_i и $p_{.j}$ (см. (2.4) и (2.4')) с помощью теоремы умножения вероятностей (см. п. 1.1.3, формулу (1.11')):

$$\begin{aligned} f_{\xi_1}(x^{(1)}, \dots, x^{(s)} \mid \xi^{(s+1)} = c^{(s+1)}, \dots, \xi^{(p)} = c^{(p)}) \\ = \frac{f_{\xi}(x^{(1)}, \dots, x^{(s)}, c^{(s+1)}, \dots, c^{(p)})}{f_{\xi_2}(c^{(s+1)}, c^{(s+2)}, \dots, c^{(p)})}. \end{aligned} \quad (2.13)$$

Аналогично

$$f_{\xi_2}(x^{(s+1)}, \dots, x^{(p)} \mid \xi^{(1)} = c^{(1)}, \dots, \xi^{(s)} = c^{(s)}) = \frac{f_{\xi}(c^{(1)}, \dots, c^{(s)}, x^{(s+1)}, \dots, x^{(p)})}{f_{\xi_1}(c^{(1)}, \dots, c^{(s)})} \quad (2.13')$$

В знаменателях правых частей (2.13) и (2.13') стоят частные (маргинальные) плотности подвекторов — соответственно $\xi_2 = (\xi^{(s+1)}, \dots, \xi^{(p)})$ и $\xi_1 = (\xi^{(1)}, \dots, \xi^{(s)})$, вычисленные в соответствии с (2.12).

Обратим внимание на существенное отличие *частной* (маргинальной) плотности $f_{\xi_1}(x^{(1)}, \dots, x^{(s)})$ от *условной* $f_{\xi_1}(x^{(1)}, \dots, x^{(s)} \mid \xi^{(s+1)} = c^{(s+1)}, \dots, \xi^{(p)} = c^{(p)})$, хотя обе они описывают распределение одного и того же набора признаков $\xi^{(1)}, \dots, \xi^{(s)}$: первая плотность не зависит от того, какие значения имеют остальные компоненты анализируемой многомерной случайной величины, в то время как *условная* плотность существенно зависит от того, на каких именно уровнях $c^{(s+1)}, \dots, c^{(p)}$ зафиксированы значения остальных компонент $\xi^{(s+1)}, \dots, \xi^{(p)}$.

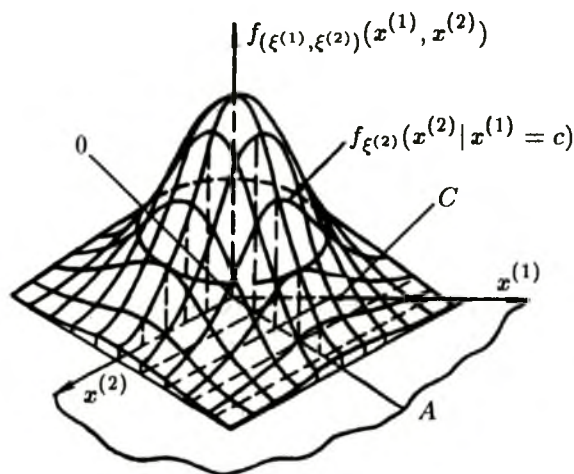


Рис. 2.6. Поверхность двумерной плотности вероятности (нормальный закон с «отсеченными» краями)

На рис. 2.6 изображен график функции плотности двумерного закона распределения вероятностей — *двумерного нормального закона* (его описание см. в п. 3.1). Там же изображены сечения поверхности двумерной плотности плоскостями $x^{(1)} = c$, т.е. плоскостями, перпендикулярными

оси $Ox^{(1)}$. В сечении получаются с точностью до нормирующего множителя одномерные законы (один из них указан стрелкой), характеризующие условное распределение компоненты $\xi^{(2)}$ при условии $\xi^{(1)} = c$. Прямая OA прослеживает характер изменения наиболее вероятного значения случайной величины $\xi^{(2)}$ в условном распределении этого признака (при условии $\xi^{(1)} = c$) в зависимости от зафиксированного значения c .

Статистическая независимость случайных величин $\xi_1, \xi_2, \dots, \xi_k$ (где признаки ξ_j могут быть дискретными и непрерывными, скалярными и векторными) вводится на базе понятия независимости системы событий (см. (1.12) и (1.13)). Случайные величины $\xi_1, \xi_2, \dots, \xi_k$ называют статистически независимыми, если для любых (измеримых) областей из возможных значений, соответственно $A_1, A_2, \dots, A_k$, имеют место соотношения

$$P\{\xi_1 \in A_1, \xi_2 \in A_2, \dots, \xi_k \in A_k\} = P\{\xi_1 \in A_1\} \dots P\{\xi_k \in A_k\}. \quad (2.14)$$

В терминах вероятностей $p_{i_1 i_2 \dots i_k}$ (для дискретных случайных величин) и плотностей $f_{(\xi_1, \dots, \xi_k)}(X^{(1)}, \dots, X^{(k)})$ (для непрерывных случайных величин) условие (2.14) может быть записано в виде:

$$\begin{aligned} p_{i_1 i_2 \dots i_k} &= P\{\xi_1 = X_{i_1}^{(1)0}, \xi_2 = X_{i_2}^{(2)0}, \dots, \xi_k = X_{i_k}^{(k)0}\} \\ &= P\{\xi_1 = X_{i_1}^{(1)0}\} P\{\xi_2 = X_{i_2}^{(2)0}\} \dots P\{\xi_k = X_{i_k}^{(k)0}\}; \end{aligned} \quad (2.14')$$

$$f_{(\xi_1, \dots, \xi_k)}(X^{(1)}, \dots, X^{(k)}) = f_{\xi_1}(X^{(1)}) \cdot f_{\xi_2}(X^{(2)}) \cdot \dots \cdot f_{\xi_k}(X^{(k)}). \quad (2.14'')$$

2.6. Основные числовые характеристики случайных величин

Итак, *исчерпывающие* сведения об интересующем нас законе распределения вероятностей можно задать и в виде полигона вероятностей (в дискретном случае), и в виде функции плотности вероятностей (в общем случае). Однако при практическом (статистическом) изучении поведения случайных величин *зачастую оказывается достаточной гораздо более скромная информация* в виде нескольких *числовых* характеристик распределения, позволяющих оценить такие его свойства, как центр группирования значений исследуемой случайной величины, мера их случайного рассеивания, степень взаимозависимости различных компонент изучаемого многомерного признака. Так, например, при изучении закона распределения заработной платы работников интересуются в первую очередь *средней*

заработной платой и одной из мер ее случайного рассеивания — *коэффициентом дифференциации* или *дисперсией*. К тому же подавляющее большинство используемых в статистических приложениях модельных законов распределения (биномиальный, пуассоновский, Парето, нормальный, логарифмически-нормальный, экспоненциальный и др., см. гл. 3) может быть однозначно восстановлено по одной-двум своим числовым характеристикам, например по среднему значению и дисперсии.

2.6.1. Понятие о математических ожиданиях и моментах

Будем рассматривать различные функции $g(\xi)$ от исследуемой случайной величины $\xi = \xi(\omega)$ (в эту схему, очевидно, может быть включен и частный случай $g(\xi) \equiv \xi$). Очевидно, и функция $g(\xi(\omega))$ будет случайной величиной, так как она является в конечном счете функцией, определенной на множестве элементарных событий ω . Результат операции «осреднения» случайной величины $g(\xi)$, произведенной с учетом «взвешивания», отвечающего заданному распределению вероятностей случайной величины ξ , носит название *математического ожидания* $g(\xi)$ и обозначается $Eg(\xi)$ ¹. Итак: если ξ — непрерывная (может быть, и многомерная) случайная величина с плотностью (совместной) вероятности $f_\xi(X)$, то

$$Eg(\xi) = \int g(X) f_\xi(X) dX \quad (2.15)$$

(интегрирование производится по области всех возможных значений признака ξ); если ξ — дискретная случайная величина с возможными значениями X_i^0 и вероятностями их осуществления p_i ($i = 1, 2, \dots$), то

$$Eg(\xi) = \sum_i g(X_i^0) \cdot p_i. \quad (2.15')$$

Важную роль в теории и практике статистических исследований играют функции $g(\xi)$ *некоторого специального вида*: $g_k(\xi) = \xi^k$ и $g_k^0(\xi) =$

¹ Если $g(\xi)$ — *векторная* функция, т. е. $g(\xi) = (g^{(1)}(\xi), g^{(2)}(\xi), \dots, g^{(p)}(\xi))$, то речь идет о покомпонентном взвешенном *осреднении*. Общепринятая символика с использованием буквы *Е* объясняется тем, что с этой буквы начинается латинское написание слова «ожидание».

$(\xi - E\xi)^k$ ($k = 1, 2, \dots$).¹ Математические ожидания функции $g_k(\xi)$ и $g_k^0(\xi)$ носят названия соответственно *начальных и центральных моментов k -го порядка* случайной величины ξ .

Рассмотрим вначале начальные и центральные моменты *одномерной* случайной величины ξ .

Итак, если для некоторого целого положительного числа k функция x^k интегрируема с весами $f_\xi(x)$ (суммируема с весами p_i) на области возможных значений ξ , то величина

$$m_k = E\xi^k = \begin{cases} \int x^k f_\xi(x) dx, & \text{если } \xi \text{ непрерывна;} \\ \sum_i (x_i^0)^k \cdot p_i, & \text{если } \xi \text{ дискретна} \end{cases} \quad (2.16)$$

называется *начальным моментом k -го порядка* или просто *k -м моментом* этого распределения или соответствующей ему случайной величины ξ , и мы говорим, что момент конечен или существует.

Очевидно, что если существует момент m_k , то существует и *центральный* момент

$$m_k^{(0)} = E(\xi - m_1)^k = \begin{cases} \int (x - m_1)^k f_\xi(x) dx, & \text{если } \xi \text{ непрерывна;} \\ \sum_i (x_i^0 - m_1)^k p_i, & \text{если } \xi \text{ дискретна.} \end{cases} \quad (2.17)$$

Раскрывая $(x - m_1)^k$ под знаком интеграла (или суммы), легко установить связи, существующие между центральными и начальными моментами:

$$m_1^{(0)} = 0;$$

¹ Если случайная величина ξ многомерна, т.е. $\xi = (\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(p)})$, то под $g_k(\xi) = \xi^k$ и $g_k^0(\xi) = (\xi - E\xi)^k$ мы будем понимать *векторные* случайные величины, компоненты которых имеют вид соответственно

$$(\xi^{(i_1)})^{k_1} \cdot (\xi^{(i_2)})^{k_2} \cdot \dots \cdot (\xi^{(i_N)})^{k_N} \text{ и} \\ (\xi^{(i_1)} - E\xi^{(i_1)})^{k_1} \cdot (\xi^{(i_2)} - E\xi^{(i_2)})^{k_2} \cdot \dots \cdot (\xi^{(i_N)} - E\xi^{(i_N)})^{k_N},$$

где $i_1 = 1, 2, \dots, p$ и $k_1 + k_2 + \dots + k_N = k$. Так, например, при $k = 2$ в качестве компонент функции $g_2(\xi) = \xi^2$ могут рассматриваться всевозможные произведения $\xi^{(i)} \cdot \xi^{(j)}$, $i, j = 1, 2, \dots, p$, в том числе, конечно, квадраты $(\xi^{(1)})^2, (\xi^{(2)})^2, \dots, (\xi^{(p)})^2$, а в качестве компонент функции $g_2^0(\xi)$ — всевозможные попарные произведения вида $(\xi^{(i)} - E\xi^{(i)})(\xi^{(j)} - E\xi^{(j)})$.

² Ниже (см. пп. 3.1.9, 3.1.10, 3.2.2, 3.2.3) будут приведены примеры ситуаций, когда моменты определенных порядков анализируемой случайной величины не существуют, т.е. операции интегрирования (суммирования) вида (2.16) приводят к «бесконечно большим» величинам.

$$\begin{aligned}
m_2^{(0)} &= m_2 - m_1^2; \\
m_3^{(0)} &= m_3 - 3m_1m_2 + 2m_1^3; \\
m_4^{(0)} &= m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4 \\
&\dots\dots\dots
\end{aligned}
\tag{2.18}$$

(ограничиваемся здесь первыми четырьмя моментами).

Наконец, при исследовании поведения *многомерных* случайных величин $\xi = (\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(p)})$ важную роль играют p^2 -мерные *векторные* функции $g(\xi) = (g^{(1)}(\xi), \dots, g^{(p^2)}(\xi))$, компонентами которых являются всевозможные попарные произведения центрированных компонент вектора ξ , т. е. элементы матрицы

$$Q = (q_{ij})_{i,j=\overline{1,p}},$$

где

$$q_{ij} = (\xi^{(i)} - m_1^{(i)})(\xi^{(j)} - m_1^{(j)}).$$

Математические ожидания элементов q_{ij} принято называть *смешанными вторыми центральными моментами* или *ковариациями* многомерного признака ξ , а матрицу

$$\Sigma = (E q_{ij})_{i,j=\overline{1,p}}, \tag{2.19}$$

составленную из ковариаций

$$\text{cov}(\xi^{(i)}, \xi^{(j)}) = E q_{ij} = E\{(\xi^{(i)} - m_1^{(i)})(\xi^{(j)} - m_1^{(j)})\} = \sigma_{ij}, \tag{2.20}$$

ковариационной матрицей признака ξ . По определению, все ковариационные матрицы являются *симметричными* (т. е. всегда $\sigma_{ij} = \sigma_{ji}$); нетрудно показать, что они являются и *неотрицательно-определенными*. Действительно, по определению (см. приложение 2) квадратная $(p \times p)$ -матрица A называется неотрицательно-определенной, если для *любого* вектора с действительными компонентами $\lambda = (\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(p)})^T$ произведение $\lambda^T \cdot A \cdot \lambda$ будет неотрицательным. Беря последовательность любых действительных чисел $\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(p)}$ и используя тождества

$$E \left[\sum_{i=1}^p \lambda^{(i)} (\xi^{(i)} - m_1^{(i)}) \right]^2 = \sum_{i=1}^p \sum_{j=1}^p \lambda^{(i)} \cdot \lambda^{(j)} \cdot \sigma_{ij} = \lambda^T \cdot \Sigma \cdot \lambda, \tag{2.21}$$

получаем доказательство неотрицательной определенности ковариационной матрицы $\Sigma = (\sigma_{ij})$, т. к. левая часть соотношения (2.21), очевидно, не может выражаться отрицательным числом.

2.6.2. Характеристики центра группирования значений случайной величины

В качестве характеристик центра группирования значений исследуемого признака в статистической практике используют несколько видов средних значений, моду и медиану. Опишем эти числовые характеристики.

Теоретическое среднее $E\xi$ исследуемой случайной величины ξ определяется как ее первый начальный момент, или, что то же, как ее *математическое ожидание* (см. (2.15) и (2.16)):

$$E\xi = m_1 = \begin{cases} \int x \cdot f_\xi(x) dx, & \text{если } \xi \text{ непрерывна;} \\ \sum_i x_i^0 \cdot p_i, & \text{если } \xi \text{ дискретна.} \end{cases} \quad (2.22)$$

Среднее значение $E\xi$ является, пожалуй, основной и наиболее употребительной характеристикой центра группирования значений случайной величины.

Непосредственно из определения среднего значения легко получить следующие его основные свойства:

- а) $Ec = c$, где c — любая неслучайная величина;
- б) $E(c\xi) = c \cdot E\xi$;
- в) $E(\xi_1 + \xi_2 + \dots + \xi_N) = E\xi_1 + E\xi_2 + \dots + E\xi_N$;
- г) $E(\xi \cdot \eta) = E\xi \cdot E\eta$, если случайные величины ξ и η статистически взаимно независимы (см. п. 2.5.3).

Среднее геометрическое (теоретическое) значение $G(\xi)$ случайной величины ξ определяется (для признаков с положительными возможными значениями) с помощью формулы

$$G(\xi) = e^{E(\ln \xi)},$$

где $e \approx 2,71828\dots$ — основание, а $\ln$ — обозначение натурального логарифма.

Можно показать, что геометрическое среднее значение $G(\xi)$ всегда меньше теоретического среднего $E\xi$.

Геометрическое среднее находит применение при расчетах темпов изменения величин и, в частности, в тех случаях, когда имеют дело с величиной, изменения которой происходят приблизительно в прямо пропорциональной зависимости с достигнутым к этому моменту уровнем самой величины (например, численность населения), или же когда имеют дело со средней из отношений, например, при расчетах «индексов цен».

Среднее гармоническое значение $H(\xi)$ случайной величины ξ задается (лишь для признаков с положительными возможными значениями) соотношением

$$H(\xi) = \frac{1}{E\left(\frac{1}{\xi}\right)}.$$

Гармоническое среднее значение ряда чисел всегда меньше геометрического среднего значения тех же чисел, а тем более — их среднего арифметического. Область его применения весьма ограничена. В экономике, в частности, пользуются иногда гармоническим средним при анализе средних норм времени, а также в некоторых видах индексных расчетов.

Модальное значение (или просто мода) x_{mod} случайной величины определяется как такое возможное значение исследуемого признака, при котором значение плотности $f_{\xi}(x)$ (в непрерывном случае) или вероятности $p\{\xi = x\}$ (в дискретном случае) достигает своего максимума. Таким образом, мода представляет собой как бы наиболее часто осуществляющееся (в экспериментах или наблюдениях), наиболее типичное значение случайной величины, т. е. значение, которое действительно является «модным»¹.

Медиана x_{med} исследуемого признака определяется как его средневероятное значение, т. е. такое значение, которое обладает следующим свойством: вероятность того, что анализируемая случайная величина окажется больше x_{med} , равна вероятности того, что она окажется меньше x_{med} . Для обладающих непрерывной плотностью непрерывных случайных величин, очевидно,

$$P\{\xi > x_{\text{med}}\} = P\{\xi < x_{\text{med}}\} = 0,5$$

и медиану можно определить как такое значение x_{med} на оси возможных значений (оси абсцисс), при котором прямая, параллельная оси ординат и проходящая через точку x_{med} , делит площадь под кривой плотности на две равные части (рис. 2.7). В некоторых случаях дискретных распределений может не существовать величины, точно удовлетворяющей сформулированному требованию. Поэтому для дискретных случайных величин

¹ Мода является естественной характеристикой центра группирования значений случайной величины лишь в случаях так называемых одновершинных (одномодальных) распределений. Многовершинные (многомодальные) распределения свидетельствуют о существенной неоднородности исследуемой совокупности. Их изучение представляет интерес в первую очередь с точки зрения задач классификации объектов и наблюдений (см. гл. 12).

медиану можно определить как *любое* число x_{med} , лежащее между двумя соседними возможными значениями $x_{i_0}^0$ и $x_{i_0+1}^0$, такими, что

$$F_{\xi}(x_{i_0}^0) < 0,5, \text{ но } F_{\xi}(x_{i_0+1}^0) \geq 0,5.$$

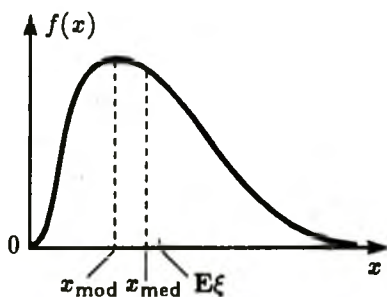


Рис. 2.7. Характер расположения моды x_{mod} , медианы x_{med} и среднего значения $E\xi$ для некоторой асимметричной плотности распределения $f(x)$

В случае симметричной плотности (или полигона распределения) среднее $E\xi$, мода x_{mod} и медиана x_{med} совпадают между собой. Для асимметричных распределений это не так (см. рис. 2.7).

2.6.3. Характеристики степени рассеяния значений случайной величины

Каждая из описанных ниже характеристик степени рассеяния — дисперсия, среднеквадратическое отклонение и коэффициент вариации — дает представление о том, как сильно могут отклоняться от своего центра группирования значения исследуемой случайной величины. Если говорить о форме кривой плотности, то эти характеристики описывают степень ее «размазанности» по всему диапазону изменения ξ : чем больше величины каждой из этих характеристик, тем более «размазанным» выглядит соответствующее распределение (рис. 2.8).

Дисперсия $D\xi$ случайной величины ξ определяется как ее второй центральный момент, т. е.

$$D\xi = m_2^{(0)} = \begin{cases} \int_x (x - E\xi)^2 f_{\xi}(x) dx, & \text{если } \xi \text{ непрерывна;} \\ \sum_i (x_i^0 - E\xi)^2 p_i, & \text{если } \xi \text{ дискретна.} \end{cases} \quad (2.23)$$

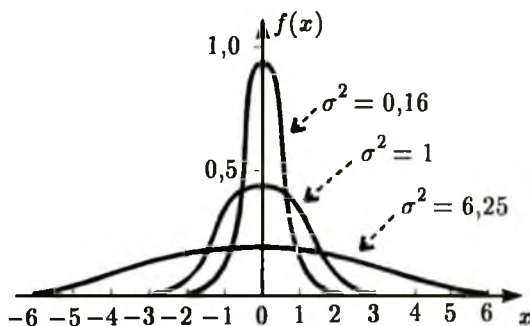


Рис. 2.8. График плотностей (нормальных законов распределения) с различными значениями дисперсии и с одинаковым (нулевым) средним значением

Из определения дисперсии (и из свойств математического ожидания) можно вывести следующие ее свойства:

- а) $Dc = 0$ (c — некоторая неслучайная величина);
- б) $D(c\xi) = c^2 D\xi$;
- в) $D(a + b\xi) = b^2 \cdot D\xi$ (a и b — некоторые неслучайные величины);
- г) $D(\xi + \eta) = D\xi + D\eta$ — в случае, когда ξ и η являются взаимно независимыми.

Часто для обозначения дисперсии используют греческую букву «сигма» (в квадрате), т. е. записывают $\sigma_\xi^2 = D\xi$.

Среднеквадратическое отклонение σ_ξ получается из дисперсии извлечением квадратного корня $\sigma_\xi = \sqrt{D\xi}$. Оно используется наряду с дисперсией для характеристики степени рассеивания случайной величины и оказывается в ряде случаев более удобным и естественным, в первую очередь, из-за своей однородности (в смысле единиц измерения) с различными характеристиками центра группирования.

Коэффициент вариации V_ξ используется в тех случаях, когда степень рассеивания естественнее описывать некоторой *относительной* характеристикой в сопоставлении со средним. В частности,

$$V_\xi = \frac{\sqrt{D\xi}}{E\xi} \cdot 100\% = \frac{\sigma_\xi}{E\xi} \cdot 100\%,$$

т. е. коэффициент вариации — это отношение (в процентах) среднеквадратического отклонения к соответствующему математическому ожиданию. Из определения ясно, что V_ξ — величина безразмерная.

2.6.4. Квантили и процентные точки распределения

При использовании различных методов математической статистики, особенно разнообразных статистических критериев (см. гл. 8) и методов построения интервальных оценок неизвестных параметров (см. гл. 7), широко используются понятия q -квантилей $u_q(F)$ и $100q$ -процентных точек $\omega_q(F)$ распределения $F(x)$.

Квантилем уровня q (или q -квантилем) непрерывной случайной величины ξ , обладающей непрерывной функцией распределения $F(x)$, называется такое возможное значение $u_q(F)$ этой случайной величины, для которого вероятность события $\xi < u_q(F)$ равна заданной величине q , т. е.

$$F(u_q) = P\{\xi < u_q\} = q. \quad (2.24)$$

Очевидно, чем больше заданное значение q ($0 < q < 1$), тем больше будет и соответствующая величина квантиля u_q . Частным случаем квантиля — 0,5-квантилем является характеристика центра группирования — медиана.

Для всякой дискретной случайной величины функция распределения $F_\xi(x)$ с увеличением аргумента x меняется, как мы видели, скачками, и, следовательно, существуют такие значения уровней q , для каждого из которых не найдется возможного значения u_q , точно удовлетворяющего уравнению (2.24). Поэтому в дискретном случае q -квантиль определяется как любое число $u_q(F)$, лежащее между двумя возможными соседними значениями $x_{i(q)}^0$ и $x_{i(q)+1}^0$, такое, что $F(x_{i(q)}^0) < q$, но $F(x_{i(q)+1}^0) \geq q$.

Часто вместо понятия квантиля используют тесно связанное с ним понятие процентной точки. Под $100q\%$ -ной точкой случайной величины ξ понимается такое ее возможное значение ω_q , для которого вероятность события $\xi \geq \omega_q$ равна q , т. е.

$$1 - F(\omega_q) = P\{\xi \geq \omega_q\} = q.$$

Для дискретных случайных величин это определение корректируется аналогично тому, как это делалось при определении квантилей.

Из определения квантилей и процентных точек вытекает простое соотношение, их связывающее:

$$u_q = \omega_{(1-q)}. \quad (2.25)$$

Для ряда наиболее часто встречающихся в статистической практике законов распределения (см. гл. 3) составлены специальные таблицы квантилей и процентных точек. Очевидно, достаточно иметь только одну из таких

таблиц, так как если, например, по таблицам процентных точек требуется найти 0,9-квантиль нормального распределения, то следует искать, в соответствии с (2.25), 10-процентную точку того же распределения.

Наглядное геометрическое представление о смысле введенных понятий дает рис. 2.9: на этом рисунке $q_1 = \int_{-\infty}^{u_{q_1}} f_{\xi}(x) dx$; $q_2 = \int_{\omega_{q_2}}^{\infty} f_{\xi}(x) dx$, а по оси Oy откладываются значения функции плотности $f_{\xi}(x)$.

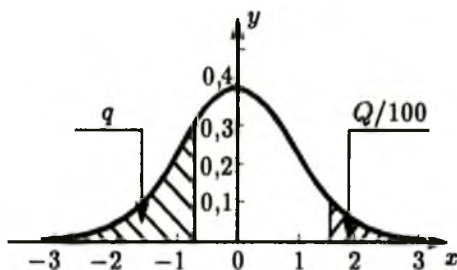


Рис. 2.9. Геометрическое пояснение смысла q_1 -квантиля u_{q_1} и $100q_2$ -процентной точки ω_{q_2} : случай стандартного нормального распределения, $q_1 = 0,25$ (соответственно $u_{0,25} = -0,675$) и $q_2 = 0,05$ (соответственно $\omega_{0,05} = 1,65$)

Квантильные характеристики помимо своей основной роли вспомогательного теоретического статистического инструментария порой играют самостоятельную роль основных характеристик изучаемого закона распределения или *содержательно интерпретируемых* параметров модели. Так, широко распространенной характеристикой степени случайного рассеяния при изучении законов распределения заработной платы и доходов являются так называемые *квантильные (уровня q) коэффициенты дифференциации $K_d(q)$* , которые определяются соотношением

$$K_d(q) = \frac{u_{1-q}}{u_q} \quad (0 < q \leq 0,25)$$

(наиболее распространенными среди них являются *децильные* коэффициенты дифференциации, когда полагают $q = 0,1$). При анализе модельных законов распределения квантили и процентные точки используют также для обозначения *практических* границ диапазона изменения исследуемого признака: так, например, квантилями уровня 0,005 и 0,995 иногда определяют соответственно минимальный и максимальный уровни заработной платы работников в соответствующей системе показателей.

2.6.5. Асимметрия и эксцесс

Обращаясь к формуле (2.17), определяющей центральные моменты распределения, легко понять, что если плотность $f_\xi(x)$ (или последовательность вероятностей $P\{\xi = x_i^0\}$) симметрична относительно среднего значения $m_1 = E\xi$ (т. е. $f(m_1 - x) \equiv f(m_1 + x)$), то все нечетные центральные моменты (если они существуют) $m_{2k+1}^{(0)}$ равны нулю. Поэтому любой нечетный, не равный нулю, момент можно рассматривать как характеристику *асимметрии* соответствующего распределения. Простейшая из этих характеристик $m_3^{(0)}$ и взята за основу вычисления так называемого **коэффициента асимметрии** β_1 — количественной характеристики *степени скошенности* распределения

$$\beta_1 = \frac{m_3^{(0)}}{(m_2^{(0)})^{\frac{3}{2}}} = \frac{E(\xi - m_1)^3}{[E(\xi - m_1)^2]^{\frac{3}{2}}}. \quad (2.26)$$

Нормировка с помощью деления $m_3^{(0)}$ на $(m_2^{(0)})^{\frac{3}{2}}$ введена для того, чтобы эта характеристика не зависела от выбора физических единиц измерения исследуемой случайной величины: формула (2.26) определяет *безразмерную* характеристику степени скошенности распределения, инвариантную относительно физических единиц измерения ξ .

Таким образом, все симметричные распределения будут иметь нулевой коэффициент асимметрии (см. рис. 2.5, 2.8, 2.9), в то время как распределения вероятностей с «длинной частью» кривой плотности, расположенной справа от ее вершины, характеризуются *положительной асимметрией* (см. рис. 2.7), а распределения с «длинной частью» кривой плотности, расположенной слева от ее вершины, обладают *отрицательной асимметрией*.

Поведение плотности (полигона) распределения в районе его модального значения обуславливает геометрическую форму соответствующей кривой в окрестности точки ее максимума, ее *островершинность*. Количественная характеристика островершинности — **эксцесс** (или **коэффициент эксцесса**) β_2 оказывается полезной характеристикой при решении ряда задач, например при определении общего вида исследуемого распределения или при его аппроксимации с помощью некоторых специальных разложений. Эта характеристика задается с помощью соотношения

$$\beta_2 = \frac{m_4^{(0)}}{(m_2^{(0)})^2} - 3 = \frac{E(\xi - m_1)^4}{[E(\xi - m_1)^2]^2} - 3. \quad (2.27)$$

Ниже мы увидим, что своеобразным началом отсчета в измерении степени островершинности служит нормальное (гауссовское) распределение,

для которого $\beta_2 = 0$.

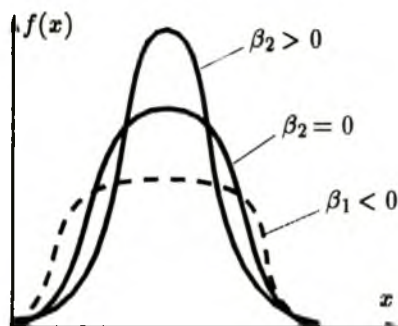


Рис. 2.10. Примеры плотностей для положительной, отрицательной и нулевой характеристик острровершинности (эксцесса)

Как правило, распределения с более высокой и более острой вершиной кривой плотности (полигона) имеют положительный эксцесс, а с менее острой — отрицательный (рис. 2.10)¹.

2.6.6. Основные характеристики многомерных распределений (ковариации, корреляции, обобщенная дисперсия и др.)

Если при описании поведения одномерных и в какой-то мере двумерных случайных величин исследователь еще имеет практически реализуемые возможности использования подходящих модельных законов распределения (см. гл. 3), то при исследовании признаков $\xi = (\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(p)})^T$ размерности большей, чем два ($p > 2$), как правило, приходится ограничиваться лишь той информацией, которую ему доставляет знание *первых двух моментов: вектором средних значений*

$$M_1 = \begin{pmatrix} m_1^{(1)} \\ m_1^{(2)} \\ \vdots \\ m_1^{(p)} \end{pmatrix} = \begin{pmatrix} E\xi^{(1)} \\ E\xi^{(2)} \\ \vdots \\ E\xi^{(p)} \end{pmatrix} \quad (2.28)$$

¹ Правда, можно построить примеры, свидетельствующие о нарушении этой общей закономерности.

и матрицей ковариаций

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \dots & \dots & \dots & \dots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix}, \quad (2.29)$$

где средние значения $m_1^{(i)}$ компонент $\xi^{(i)}$ определяются в соответствии с формулой (2.22) с использованием одномерных частных плотностей (полигонов) распределения случайных величин $\xi^{(i)}$, а ковариации $\sigma_{jk} = E\{(\xi^{(j)} - m_1^{(j)})(\xi^{(k)} - m_1^{(k)})\}$ подсчитываются с использованием соответствующих двумерных частных плотностей (полигонов) распределения пары случайных величин $(\xi^{(j)}, \xi^{(k)})$.

Как и в одномерном случае, вектор средних значений является основной характеристикой центра группирования наблюдений исследуемого многомерного признака ξ (в соответствующем p -мерном пространстве ее возможных значений).

Матрица ковариаций Σ характеризует следующие свойства исследуемого многомерного признака.

1. **Степень случайного разброса** отдельно по каждой компоненте и в целом по многомерному признаку. Легко видеть, что диагональные элементы σ_{ii} матрицы Σ определяют частные дисперсии компонент $\xi^{(i)}$, т. е. степень случайного разброса значений одномерной случайной величины $\xi^{(i)}$. Итак,

$$\sigma_{ii} = E(\xi^{(i)} - m_1^{(i)})^2 = D\xi^{(i)}.$$

Многомерным аналогом дисперсии является величина определителя ковариационной матрицы, называемая *обобщенной дисперсией* многомерного случайного признака ξ^1 :

$$D_{\text{об}}\xi = \det(\Sigma). \quad (2.30)$$

Часто используется и другая характеристика степени случайного рассеяния значений многомерной случайной величины — так называемый след ковариационной матрицы Σ , т. е. сумма ее диагональных элементов:

$$\text{tr}(\Sigma) = \sigma_{11} + \sigma_{22} + \dots + \sigma_{pp}. \quad (2.31)$$

¹ Мы придерживаемся распространенных матричных обозначений: $\det(A)$ или $|A|$ — определитель или детерминант (determinant, англ.) матрицы A ; $\text{tr}(A)$ — след (trace, англ.) матрицы A .

Из неотрицательной определенности матрицы Σ (см. п. 2.6.1) и смысла диагональных элементов σ_{ii} следует, что величины, определенные соотношениями (2.30) и (2.31), всегда неотрицательны.

Поясним *геометрический смысл* обобщенной дисперсии. Можно сказать, что если, например, исследуемый многомерный признак подчинен нормальному закону распределения (см. гл. 3), то для любого заданного уровня вероятности P_0 объем области (окружающей центр группирования M_1), вероятность попадания в которую значений анализируемой случайной величины ξ равна P_0 , пропорционален $\sqrt{D_{06}\xi}$ (этот объем пропорционален также некоторому множителю, зависящему от размерности p , и пропорционален, кроме того, некоторому числу, определяемому в зависимости от заданного уровня вероятности P_0).

Характер и структура статистических взаимосвязей, существующих между компонентами анализируемого многомерного признака, также могут быть описаны с помощью ковариационной матрицы. Однако в этом случае удобнее перейти к определенным образом *нормированной* ковариационной матрице, называемой *корреляционной*, а именно к матрице

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{pmatrix}, \quad (2.32)$$

где элементы r_{jk} получаются из элементов σ_{jk} с помощью нормировки

$$r_{jk} = \frac{\sigma_{jk}}{(\sigma_{jj}\sigma_{kk})^{1/2}}. \quad (2.33)$$

Характеристики r_{jk} называются *коэффициентами корреляции* между случайными величинами $\xi^{(j)}$ и $\xi^{(k)}$, являются *измерителями степени тесноты линейной статистической связи* между этими признаками и обладают следующими свойствами:

а) $-1 \leq r_{jk} \leq 1$, что следует непосредственно из неравенств

$$E \left(\frac{\xi^{(j)} - m_1^{(j)}}{\sigma_{jj}^{1/2}} \pm \frac{\xi^{(k)} - m_1^{(k)}}{\sigma_{kk}^{1/2}} \right)^2 \geq 0;$$

б) максимальная степень тесноты связи соответствует значениям коэффициента корреляции, равным $+1$ или -1 , и достигается либо при измерении связи признака с самим собой (тогда, очевидно, $r_{jj} \equiv 1$), либо при наличии *линейной функциональной* связи между $\xi^{(j)}$ и $\xi^{(k)}$, т. е. в случае

$\xi^{(j)} = b_0 + b_1 \xi^{(k)}$, где b_0 и b_1 — некоторые постоянные величины (при этом если $b_1 > 0$, то связь называется *положительной*, а если $b_1 < 0$, то связь называется *отрицательной*);

в) если случайные компоненты $\xi^{(j)}$ и $\xi^{(k)}$ статистически независимы, то $r_{jk} = 0$ (следует непосредственно из того факта, что для независимых случайных величин $\xi^{(j)}$ и $\xi^{(k)}$ $E(\xi^{(j)} \cdot \xi^{(k)}) = (E\xi^{(j)})(E\xi^{(k)})$). Обратное утверждение (из $r_{jk} = 0$ следует независимость $\xi^{(j)}$ и $\xi^{(k)}$) верно лишь для некоторых частных случаев (например, для нормально распределенных пар $(\xi^{(j)}, \xi^{(k)})$) и неверно в общем случае.

ВЫВОДЫ

1. *Случайная величина* (случайный признак) определяет перечень показателей — количественных, ординальных (порядковых) или номинальных (классификационных), которые подлежат статистическому исследованию в ходе проводимых случайных экспериментов (наблюдений).

2. «Возможные значения»¹ случайной величины определяются природой и составом пространства элементарных событий Ω : каждому элементарному исходу ω соответствует «возможное значение» $x(\omega)$ исследуемой случайной величины ξ , поэтому последнюю можно определить как функцию $\xi = \xi(\omega)$, заданную на множестве элементарных событий.

3. *Закон распределения вероятностей* исследуемой случайной величины позволяет сопоставлять с любой измеримой областью ΔX возможных значений вероятность $p(\Delta X)$ события, заключающегося в том, что реализовавшееся в результате случайного эксперимента (наблюдения) значение данной случайной величины окажется принадлежащим этой области, т. е.

$$p(\Delta X) = P\{\xi \in \Delta X\}.$$

4. Для описания закона распределения вероятностей *многомерного признака* $\xi = (\xi^{(1)}, \dots, \xi^{(p)})$ могут, как и в одномерном случае, использоваться функция распределения $F_\xi(x^{(1)}, \dots, x^{(p)}) = P\{\xi^{(1)} < x^{(1)}, \dots, \xi^{(p)} < x^{(p)}\}$ и функция плотности $f_\xi(x^{(1)} \dots x^{(p)}) = \partial^p F_\xi(x^{(1)}, \dots, x^{(p)}) / \partial x^{(1)}$,

¹ Помимо числовой (скалярной) формы эти «возможные значения» могут быть представлены и в векторной, и в матричной форме и, более того, могут быть определены в пространстве самой общей природы (в зависимости от сущности и целей исследуемого случайного эксперимента).

$\dots, \partial x^{(p)}$. Однако в отличие от одномерного случая исчерпывающей формой задания многомерного закона распределения является только совместная функция плотности вероятности.

5. Зная совместный закон распределения многомерной случайной величины $\xi = (\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(p)})$, можно получить частный (маржинальный) закон распределения любого подвектора $\xi_1 = (\xi^{(i_1)}, \xi^{(i_2)}, \dots, \xi^{(i_k)})$, $1 \leq i_q \leq p$, $1 \leq k < p$, а также условный закон распределения, описывающий распределение любого подвектора ξ_1 , когда все или какая-то часть остальных компонент исходного векторного признака фиксируются на заданных уровнях (см. (2.12) и (2.13)).

6. Если компоненты $\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(p)}$ анализируемого случайного признака $\xi = (\xi^{(1)}, \dots, \xi^{(p)})$ статистически независимы, то многомерный закон распределения $f_\xi(x^{(1)}, \dots, x^{(p)})$ может быть описан p частными одномерными законами, так как в этом случае по определению

$$f_\xi(x^{(1)}, x^{(2)}, \dots, x^{(p)}) = f_{\xi^{(1)}}(x^{(1)}) \cdot f_{\xi^{(2)}}(x^{(2)}) \dots f_{\xi^{(p)}}(x^{(p)}).$$

7. При практическом изучении поведения исследуемого случайного признака ξ зачастую оказывается достаточным знания ограниченного набора его числовых характеристик: среднего значения $E\xi$, дисперсии $D\xi$, коэффициентов асимметрии и эксцесса (β_1 и β_2), а в многомерном случае — еще элементов σ_{jk} ковариационной матрицы Σ .

ГЛАВА 3. МОДЕЛИ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ, НАИБОЛЕЕ РАСПРОСТРАНЕННЫЕ В ПРАКТИКЕ СТАТИСТИЧЕСКИХ ИССЛЕДОВАНИЙ

Говоря о распространенности той или иной модели распределения в практике статистических исследований, следует иметь в виду две возможные роли, которые эта модель может играть. Первая из них заключается в адекватном описании механизма исследуемого *реального процесса*, генерирующего подлежащие статистическому анализу исходные статистические данные. В этом случае выбранная по тем или иным соображениям (или выведенная теоретически) *модель описывает закон распределения вероятностей непосредственно анализируемой и имеющей четкую физическую интерпретацию случайной величины* (заработной платы работника, дохода семьи, числа сбоев автоматической линии в единицу времени, числа дефектных изделий, обнаруженных в проконтролированной партии заданного объема, и т. д.). Подходы к построению таких моделей, методы их анализа и обоснования относятся к области «реалистического» (или содержательного) моделирования.

Другая роль широко распространенных в статистических исследованиях моделей — использование их в качестве вспомогательного технического средства при реализации методов статистической обработки данных. С помощью моделей этого типа описываются распределения вероятностей некоторых вспомогательных функций от исследуемых случайных величин. Эти функции используются при построении разного рода статистических оценок и статистических критериев (о способах построения оценок и критериев см. главы 6–8). К распределениям этого типа относятся в первую очередь распределения «хи-квадрат», Стьюдента (t -распределение) и F -распределение.

Этой условной классификации распределений мы и будем придерживаться при изложении содержания данной главы.

3.1. Законы распределения, используемые для описания механизмов генерации реальных социально-экономических данных

3.1.1. Распределения, возникающие при анализе последовательности испытаний Бернулли: биномиальное и отрицательное биномиальное

Широкий класс случайных величин, которые приходится изучать в практике статистических исследований, индуцируется последовательностью независимых случайных экспериментов следующего типа: в результате реализации каждого случайного эксперимента (наблюдения) некоторое интересующее нас событие A может произойти (с некоторой вероятностью p) или не произойти (соответственно с вероятностью $q = 1 - p$); при многократном (m -кратном) повторении этого эксперимента *вероятность p осуществления события A остается одной и той же, а наблюдения, составляющие эту последовательность экспериментов, являются взаимно независимыми*. Серию экспериментов подобного типа принято называть *последовательностью испытаний Бернулли*. Можно описать эту последовательность в терминах случайных величин, сопоставляя с i -м по счету экспериментом данной последовательности случайную величину

$$\xi_i = \begin{cases} 1, & \text{если событие } A \text{ произошло;} \\ 0, & \text{если событие } A \text{ не произошло.} \end{cases} \quad (3.1)$$

Тогда «бернуллиевость» последовательности $\xi_1, \xi_2, \dots, \xi_m$ означает, что $P\{\xi_1 = 1\} = P\{\xi_2 = 1\} = \dots = P\{\xi_m = 1\} = p$, причем случайные величины $\xi_1, \xi_2, \dots, \xi_m$ статистически независимы (определение статистической независимости случайных величин дано в п.2.5.3).

При определенных (как правило, приблизительно соблюдающихся на практике) условиях в схему испытаний Бернулли хорошо укладываются такие случайные эксперименты, как бросание монеты или игральной кости, проверка (по альтернативному признаку) изделий массовой продукции, обращение к «обслуживаемому устройству» (с исходами «свободен — занят»), попытка выполнения некоторого задания (с исходами «выполнено — не выполнено»), стрельба по цели (с исходами «попадание — промах») и т. п.

«Единичное» испытание Бернулли можно интерпретировать и как извлечение объекта из воображаемой бесконечной совокупности, в которой доля p объектов обладает некоторым интересующим нас свойством. Тогда

интересующее нас событие A заключается в том, что при этом извлечении мы «вытащим» один из объектов, обладающих упомянутым свойством.

Биномиальный закон описывает распределение случайной величины $\nu_p(m) = \xi_1 + \xi_2 + \dots + \xi_m$, т.е. числа появления интересующего нас события в последовательности из m независимых испытаний, когда вероятность появления этого события в одном испытании равна p .

Из определения биномиальной случайной величины следует, что ее возможными значениями являются все целые неотрицательные числа от нуля до m . Для вывода вероятностей $P\{\nu_p(m) = x\}$ ($x = 0, 1, 2, \dots, m$) рассмотрим внимательнее пространство элементарных событий, порожденное последовательностью испытаний Бернулли. Очевидно, каждому элементарному событию ω соответствует последовательность из нулей и единиц длины m

$$\delta_1(\omega), \delta_2(\omega), \dots, \delta_m(\omega). \quad (3.2)$$

Разобьем эти последовательности на классы, включая в один (x -й) класс все последовательности типа (3.2), содержащие *одинаковое* число x единиц:

[illegible]

Имея в виду, что число $N(x)$ элементарных событий в классе с номером x равно числу сочетаний из m элементов по x (поскольку x единиц можно разместить на m местах именно таким числом различных способов), а также тот факт, что вероятность осуществления любого элементарного исхода, входящего в класс с номером x , равна, как нетрудно под-

считать, величине $p^x(1-p)^{m-x}$, получаем¹

$$\begin{aligned} P\{\nu_p(m) = x\} &= P\{\omega_1(x) + \omega_2(x) + \dots + \omega_{C_m^x}(x)\} \\ &= \sum_1^{C_m^x} p^x(1-p)^{m-x} = C_m^x p^x(1-p)^{m-x}. \end{aligned} \quad (3.3)$$

Это и есть формула (аналитическая запись, модель) биномиального

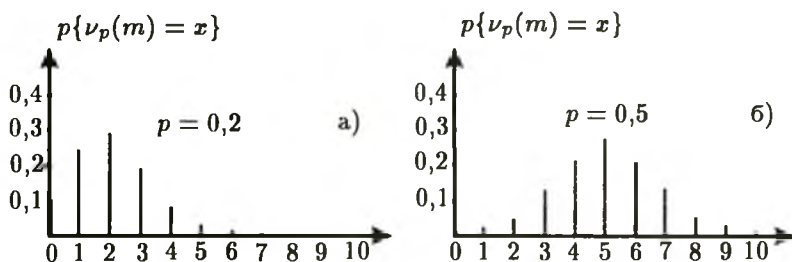


Рис. 3.1. Полигон вероятностей биномиального распределения для различных значений p и $m = 10$: а) $p = 0,2$; $m = 10$; б) $p = 0,5$; $m = 10$

закона¹ распределения (см. рис. 3.1). Подсчет его основных числовых характеристик (который в данном случае легче реализовать, не используя прямые формулы типа (2.17)–(2.18), а опираясь на соотношение $\nu_p(m) = \xi_1 + \dots + \xi_m$, взаимную независимость ξ_i и простоту вычисления их моментов) дает:

<i>среднее:</i>	$E\nu_p(m) = mp$;
<i>мода $x_{\text{мод}}$:</i>	$p(m+1) - 1 \leq x_{\text{мод}} \leq p(m+1)$;
<i>дисперсия:</i>	$D\nu_p(m) = mp(1-p)$;
<i>асимметрия:</i>	$\beta_1 = \frac{1-2p}{\sqrt{mp(1-p)}}$;
<i>эксцесс:</i>	$\beta_2 = \frac{1-6p(1-p)}{mp(1-p)}$.

¹ Далее используется одно из широко используемых обозначений для числа сочетаний из m элементов по x , а именно C_m^x .

² «Биномиальным» этот закон называется потому, что правая часть (3.3) представляет собой x -й член биномиального разложения $(p + (1-p))^m$. Этот же факт позволяет легко проверить соблюдение аксиомы о нормировке вероятностей, т. е. справедливость тождества $\sum_{x=0}^m C_m^x p^x(1-p)^{m-x} \equiv 1$.

Биномиальное распределение широко используется в теории и практике статистического контроля качества продукции, при описании функционирования систем массового обслуживания, в теории стрельбы и в других областях практической деятельности.

Отрицательный биномиальный закон описывает распределение случайной величины $\nu_p^-(k)$, определяемой испытаниями Бернулли $\xi_1, \xi_2, \dots$ (см. (3.1)) следующим образом:

$$\sum_{i=1}^{\nu_p^-(k)-1} \xi_i = k - 1, \quad \sum_{i=1}^{\nu_p^-(k)} \xi_i = k.$$

Другими словами $\nu_p^-(k)$ — это число испытаний в схеме Бернулли (с вероятностью p появления интересующего нас события в результате проведения одного испытания) *до k -го появления интересующего нас события* (включая последнее испытание). Нетрудно вывести аналитический вид распределения случайной величины $\nu_p^-(k)$. Зафиксируем любое ее возможное значение x . Из того, что при числе испытаний $\nu_p^-(k) = x$ впервые осуществилось заданное число k появлений интересующего нас события, следует, что на предыдущем шаге, т. е. при числе испытаний, равном $x-1$, мы имели $k-1$ появлений того же события. Следовательно, опираясь на теорему умножения вероятностей, мы можем выразить вероятность $P\{\nu_p^-(k) = x\}$ в терминах вероятностей, связанных с поведением биномиальных случайных величин $\nu_p(x-1)$ и $\nu_p(1)$, а именно:

$$\begin{aligned} P\{\nu_p^-(k) = x\} &= P\{\nu_p(x-1) = k-1\} \cdot P\{\nu_p(1) = 1\} \\ &= [C_{x-1}^{k-1} p^{k-1} (1-p)^{(x-1)-(k-1)}] \cdot p \\ &= C_{x-1}^{k-1} p^k (1-p)^{x-k}, \quad x = k, k+1, \dots \end{aligned} \quad (3.4)$$

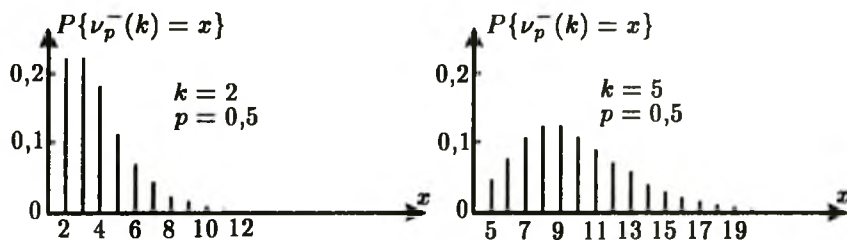


Рис. 3.2. Полигон вероятностей отрицательного биномиального закона распределения для различных значений k при $p = 0,5$

Название данного закона объясняется тем, что правые части (3.4) являются последовательными членами разложения некоторого *бинома с отрицательным показателем*. Соответствующий (3.4) полигон вероятностей изображен на рис. 3.2.

Основные числовые характеристики закона:

$$\begin{aligned}
 \text{среднее:} \quad & E\nu_p^-(k) = \frac{k}{p}; \\
 \text{дисперсия:} \quad & D\nu_p^-(k) = \frac{k(1-p)}{p^2}; \\
 \text{асимметрия:} \quad & \beta_1 = \frac{2-p}{\sqrt{k(1-p)}}; \\
 \text{эксцесс:} \quad & \beta_2 = \frac{1+4(1-p)+(1-p)^2}{k(1-p)}.
 \end{aligned}$$

Модель отрицательного биномиального распределения применяется в статистике несчастных случаев и заболеваний, в задачах, связанных с анализом количеств индивидуумов данного вида в выборках из биологических совокупностей, в задачах оптимального резервирования элементов, в теории стрельбы.

3.1.2. Гипергеометрическое распределение

В одном из вариантов интерпретации биномиальной случайной величины $\nu_p(t)$ мы рассматривали некоторую воображаемую *бесконечную* совокупность, доля p объектов которой обладает интересующим нас свойством. Тогда $\nu_p(t)$ означает число объектов, обладающих этим свойством среди t объектов, случайно извлеченных из данной совокупности. Гипергеометрическую случайную величину $\nu_{M,N}(t)$ можно считать модификацией биномиальной случайной величины $\nu_p(t)$, приспособленной к случаю *конечной* совокупности, состоящей из N объектов, среди которых имеются M объектов с интересующим нас свойством. Иначе говоря, $\nu_{M,N}(t)$ — это число объектов, обладающих заданным свойством среди t объектов, случайно извлеченных (без возвращения) из совокупности N объектов, M из которых обладают этим свойством. Очевидно, возможными значениями случайной величины $\nu_{M,N}(t)$ будут все целые неотрицательные числа от $\max\{0, t - (N - M)\}$ до $\min\{t, M\}$. Для вывода аналитического вида ее закона распределения подсчитаем вероятность события $\{\nu_{M,N}(t) = x\}$ как отношение числа всевозможных выборок объема t , приводящих к осуществлению этого события (числа «благоприятных»

исходов), к общему числу способов, которыми можно выбрать m объектов из N (к числу всех возможных исходов). Очевидно, каждому набору из x объектов с заданным свойством соответствует C_{N-M}^{m-x} способов, которыми можно отобрать остальные $m - x$ объектов из числа объектов, не обладающих этим свойством. А поскольку такие наборы из x объектов с заданным свойством можно сформировать C_M^x различными способами, то общее число «благоприятных» (для события $\{\nu_{M,N}(m) = x\}$) исходов будет $C_M^x C_{N-M}^{m-x}$. Учитывая, что число всех возможных исходов, т. е. всех возможных способов, которыми можно извлечь m объектов из N предложенных, равно C_N^m , получаем

$$P\{\nu_{M,N}(m) = x\} = \frac{C_M^x \cdot C_{N-M}^{m-x}}{C_N^m}. \quad (3.5)$$

Этот закон широко используется в практике статистического приемочного контроля качества промышленной продукции, а также в различных задачах, связанных с организацией выборочных обследований.

Его основные числовые характеристики:

среднее:
$$E\nu_{M,N}(m) = m \frac{M}{N};$$

дисперсия:
$$D\nu_{M,N}(m) = m \frac{M}{N-1} \left(1 - \frac{M}{N}\right) \left(1 - \frac{m}{N}\right);$$

асимметрия:
$$\beta_1 = \frac{(1 - 2\frac{M}{N})}{\sqrt{m\frac{M}{N}(1 - \frac{M}{N})}} \frac{(N - 2m)\sqrt{N-1}}{(N-2)\sqrt{N-m}};$$

эксцесс:
$$\beta_2 = \frac{c_1(N) - c_2(N) 6\frac{M}{N}(1 - \frac{M}{N})}{m\frac{M}{N}(1 - \frac{M}{N})} + c_3(N) + c_4(N),$$

где

$$c_1(N) = \frac{(N-1)N(N+1)}{(N-2)(N-3)(N-m)};$$

$$c_2(N) = \frac{(N-1)N^2}{(N-2)(N-3)(N-m)};$$

$$c_3(N) = 3 \left[\frac{(N-1)N^2}{(N-2)(N-3)(N-m)} - 1 \right];$$

$$c_4(N) = \frac{18(N-1)}{(N-2)(N-3)} - \frac{6(N-1)}{(N-2)(N-3)m\frac{M}{N}\left(1-\frac{M}{N}\right)} - \frac{3(N-1)Nm}{(N-2)(N-3)(N-m)}.$$

При $N \rightarrow \infty$ правая часть (3.5) стремится, как и следовало ожидать, к выражению для биномиального закона распределения (3.3), и соответственно среднее значение, дисперсия, асимметрия и эксцесс гипергеометрического распределения сходятся к аналогичным числовым характеристикам биномиально распределенной случайной величины (что легко устанавливается с помощью соответствующего предельного перехода).

3.1.3. Распределение Пуассона

Если нас интересует число наступлений определенного случайного события за единицу времени, когда факт наступления этого события в данном эксперименте не зависит от того, сколько раз и в какие моменты времени оно осуществлялось в прошлом, и не влияет на будущее, а испытания производятся в стационарных условиях, то для описания распределения такой случайной величины обычно используют *закон Пуассона* (данное распределение впервые предложено и опубликовано этим ученым в 1837 г.). Этот закон можно также описывать как *предельный случай биномиального распределения*, когда вероятность p осуществления интересующего нас события в единичном эксперименте очень мала, но число экспериментов m , производимых в единицу времени, достаточно велико, а именно такое, что в процессе $p \rightarrow 0$ и $m \rightarrow \infty$ произведение mp стремится к некоторой положительной постоянной величине λ (т. е. $mp \rightarrow \lambda$). Поэтому закон Пуассона часто называют также *законом редких событий*. Обозначим пуассоновскую случайную величину $\nu_0(\infty)$ или просто ν_0 (имея в виду предельный переход от биномиальной случайной величины $\nu_p(m)$ по $p \rightarrow 0$ и $m \rightarrow \infty$) и выведем ее закон распределения:

$$\begin{aligned} P\{\nu_0 = x\} &= \lim_{\substack{p \rightarrow 0, m \rightarrow \infty \\ (mp \rightarrow \lambda)}} P\{\nu_p(m) = x\} = \lim C_m^x p^x (1-p)^{m-x} \\ &= \lim \frac{m(m-1)\dots(m-x+1)}{x!} \frac{\lambda^x}{m^x} \left(1 - \frac{\lambda}{m}\right)^{m-x} \\ &= \lim \left[\frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{m}\right)^m \left(1 - \frac{1}{m}\right) \left(1 - \frac{2}{m}\right) \dots \right. \\ &\quad \left. \left(1 - \frac{x-1}{m}\right) \left(1 - \frac{\lambda}{m}\right)^{-x} \right] \end{aligned}$$

$$\begin{aligned}
 &= \frac{\lambda^x}{x!} \lim_{m \rightarrow \infty} \left(1 - \frac{\lambda}{m}\right)^m \lim_{m \rightarrow \infty} \left[\left(1 - \frac{1}{m}\right) \left(1 - \frac{2}{m}\right) \right. \\
 &\quad \left. \dots \left(1 - \frac{x-1}{m}\right) \left(1 - \frac{\lambda}{m}\right)^{-x} \right] \\
 &= \frac{\lambda^x}{x!} e^{-\lambda} \quad (x = 0, 1, 2, \dots). \quad (3.6)
 \end{aligned}$$

Как видим, закон распределения Пуассона зависит от *единственного* параметра λ , содержательно интерпретируемого как среднее число осуществления интересующего нас события в единицу времени (см. рис. 3.3).

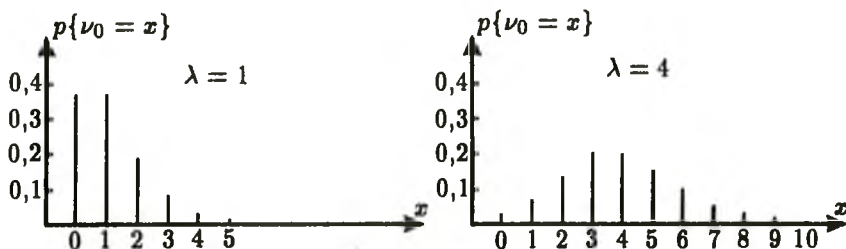


Рис. 3.3. Полигон вероятностей пуассоновского закона распределения для различных значений λ

С помощью «прямого счета» по формулам (2.17)–(2.18) могут быть подсчитаны основные числовые характеристики пуассоновской случайной величины:

<i>среднее:</i>	$E\nu_0 = \lambda;$
<i>дисперсия:</i>	$D\nu_0 = \lambda;$
<i>асимметрия:</i>	$\beta_1 = \frac{1}{\sqrt{\lambda}};$
<i>эксцесс:</i>	$\beta_2 = \frac{1}{\lambda}.$

Пуассоновская случайная величина используется для описания числа сбоев автоматической линии или числа отказов сложной системы (работающих в «нормальном» режиме) в единицу времени; числа «требований на обслуживание», поступивших в единицу времени в систему массового обслуживания; статистических закономерностей несчастных случаев и редких заболеваний.

Привлекательные прикладные свойства этого закона не исчерпываются вычислительными удобствами и лаконичностью формулы (3.6) (модель зависит всего от *одного числового параметра* λ !). Оказывается, эта

модель остается работоспособной и в ситуациях, отклоняющихся от вышеописанной схемы ее формирования. Например, можно допустить, что разные бернуллиевские испытания имеют *разные* вероятности осуществления интересующего нас события $p_1, p_2, \dots, p_m$. В этом случае биномиальный закон применительно к такой серии испытаний уже не может быть применен, в то время как выражение (3.6) остается приблизительно справедливым и дает достаточно точное описание распределения интересующей нас случайной величины, если только в него вместо $\lambda = mp$ доставить величину $\lambda = m\bar{p}$, где $\bar{p} = (p_1 + \dots + p_m)/m$. Сказанное означает, что можно предположить, что анализируемая совокупность состоит из смеси множества разнородных подсовокупностей, таких, что при переходе из одной подсовокупности в другую меняется доля p содержащихся в них объектов с заданным свойством, а следовательно, меняется и среднее число λ осуществления интересующего нас события в единицу времени. Можно далее показать, что если вместо использования среднего значения этих p (или λ) (при котором мы остаемся в рамках модели (3.6)) ввести в рассмотрение *закон распределения меняющегося* параметра λ , интерпретируемого как *случайная величина*, то мы придем к другому, но в определенном смысле близкому к пуассоновскому закону распределения. Так, например, если предположить, что функция плотности распределения случайного параметра λ имеет вид

$$f_{\lambda}(x) = \frac{p^k}{\Gamma(k)} \cdot x^{k-1} e^{-px},$$

где $\Gamma(k) = \int_0^{\infty} x^{k-1} e^{-x} dx$ — гамма-функция Эйлера, положительные числа k и p — параметры закона, а $x > 0$ — возможные значения λ , число осуществления (в единицу времени) интересующего нас события будет подчинено известному нам отрицательному биномиальному закону (3.4) (подробнее о распределении $f_{\lambda}(x)$ см. в п. 3.2.5).

3.1.4. Полиномиальное (мультиномиальное) распределение

Полезным обобщением биномиального распределения на случай *более чем двух* возможных исходов является полиномиальный (мультиномиальный) закон распределения. Можно интерпретировать анализируемую в этом случае ситуацию таким образом, что мы имеем дело с некоторой бесконечной совокупностью, содержащей объекты l различных типов ($l \geq 2$), представленных соответственно в долях $p_1, p_2, \dots, p_l$ (в биномиальной генеральной совокупности мы имели $l = 2$, $p_1 = p$ и $p_2 = 1 - p$). Таким образом, в результате одного случайного эксперимента (случай-

ного извлечения объекта из этой бесконечной совокупности) объект типа j появляется с вероятностью p_j . Нас будет интересовать распределение многомерной случайной величины $(\nu_p^{(1)}(m), \nu_p^{(2)}(m), \dots, \nu_p^{(l)}(m))$, порожденной m -кратным случайным экспериментом (т.е. извлечением m объектов), где $\nu_p^{(j)}(m)$ — число объектов j -го типа, оказавшихся в этой выборке, а $p = (p_1, p_2, \dots, p_l)$ (очевидно $\sum_{j=1}^l p_j = 1$ и $\sum_{j=1}^l \nu_p^{(j)}(m) = m$).

Соответствующее многомерное дискретное распределение описывается выражением (доказывается прямым вероятностным рассуждением)

$$\begin{aligned} P\{\nu_p^{(1)}(m) = x^{(1)}, \dots, \nu_p^{(l)}(m) = x^{(l)}\} \\ = \frac{m!}{x^{(1)}! x^{(2)}! \dots x^{(l)}!} p_1^{x^{(1)}} \dots p_l^{x^{(l)}}, \end{aligned} \quad (3.7)$$

где $x^{(1)}, x^{(2)}, \dots, x^{(l)}$ — любые (заданные) целые неотрицательные числа, подчиненные условию $\sum_{j=1}^l x^{(j)} = m$, а выражение (3.7) определяет вероятность того, что среди m извлеченных объектов оказалось ровно $x^{(1)}$ объектов 1-го типа, $x^{(2)}$ объектов 2-го типа и т.д. Можно также связывать полиномиальную случайную величину с m -кратным случайным экспериментом, каждый из которых может закончиться одним из l возможных исходов $A_1, A_2, \dots, A_l$, причем вероятность исхода A_j в единичном эксперименте равна p_j .

Название распределения объясняется тем, что выражение (3.7) является общим членом разложения многочлена (полинома) $(p_1 + p_2 + \dots + p_l)^m$.

Вектор средних значений $(E\nu_p^{(1)}(m), \dots, E\nu_p^{(l)}(m))$ и ковариации $\sigma_{jk} = E\{(\nu_p^{(j)}(m) - E\nu_p^{(j)}(m))(\nu_p^{(k)}(m) - E\nu_p^{(k)}(m))\}$ компонент исследуемой многомерной случайной величины определяются соотношениями:

$$\begin{aligned} \text{средние:} \quad & E\nu_p^{(j)}(m) = mp_j; \quad j = 1, 2, \dots, l; \\ \text{дисперсии:} \quad & D\nu_p^{(j)}(m) = \sigma_{jj} = mp_j(1 - p_j), \quad j = 1, 2, \dots, l; \\ \text{ковариации:} \quad & \sigma_{jk} = -mp_j p_k; \quad j, k = 1, 2, \dots, l, \quad j \neq k. \end{aligned}$$

Полиномиальное распределение применяется главным образом при статистической обработке выборок из больших совокупностей, элементы которых разделяются более чем на две категории (например, в различных социологических, экономико-социологических, медицинских и других выборочных обследованиях).

3.1.5. Нормальное (гауссовское) распределение

Это распределение занимает центральное место в теории и практике вероятностно-статистических исследований. В качестве непрерывной аппроксимации к биномиальному распределению оно впервые рассматривалось А. Муавром еще в 1733 г. (см. ниже теорему Муавра–Лапласа, п. 4.3). Некоторое время спустя нормальное распределение было снова открыто и изучено независимо друг от друга К. Гауссом (1809 г.) и П. Лапласом (1812 г.). Оба ученых пришли к нормальному закону в связи со своей работой по теории ошибок наблюдений. Идея их объяснения механизма формирования нормально распределенных случайных величин заключается в следующем. Постулируется, что значения исследуемой непрерывной случайной величины формируются *под воздействием очень большого числа независимых случайных факторов, причем сила воздействия каждого отдельного фактора мала и не может превалировать среди остальных, а характер воздействия — аддитивный* (т. е. при воздействии случайного фактора F на величину a получается величина $a + \Delta(F)$, где случайная «добавка» $\Delta(F)$ мала и равновероятна по знаку¹. Можно показать, что функция плотности распределения случайных величин подобного типа имеет вид

$$\varphi(x; a, \sigma^2) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad (3.8)$$

где a и σ^2 — параметры закона, интерпретируемые соответственно как среднее значение и дисперсия данной случайной величины (в виду особой роли нормального распределения мы будем использовать специальную символику для обозначения его функции плотности и функции распределения).

Соответствующая функция распределения нормальной случайной величины $\xi(a, \sigma^2)$ обозначается $\Phi(x; a, \sigma^2)$ и задается соотношением

$$\Phi(x; a, \sigma^2) = P\{\xi(a, \sigma^2) < x\} = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(t-a)^2}{2\sigma^2}} dt. \quad (3.8')$$

Условимся называть нормальный закон с параметрами $a = 0$ и $\sigma^2 = 1$ *стандартным*, а его функции плотности и распределения обозначать соответственно $\varphi(x) = \varphi(x; 0, 1)$ и $\Phi(x) = \Phi(x; 0, 1)$. Соответственно

¹ Строгая теоретическая формализация этих условий содержится, например, в центральной предельной теореме, см. п. 4.3.

нормальный з.р.в. с произвольными значениями параметров a и σ^2 будем называть (a, σ^2) нормальным.

Во многих случайных величинах, изучаемых в экономике, технике, медицине, биологии и в других областях, естественно видеть суммарный аддитивный эффект большого числа независимых причин. Но центральное место нормального закона не следует объяснять его универсальной приложимостью, как это было принято долгое время (по-видимому, под влиянием выдающихся работ К. Гаусса и П. Лапласа).

В этом смысле нормальный закон — это один из многих типов распределения, имеющих в природе, правда, с относительно большим удельным весом практической приложимости. И потому нам понятна ирония, звучащая в известном высказывании Липмана (цитируемом А. Пуанкаре в своем труде «Исчисление вероятностей», Париж, 1912 г.): «Каждый уверен в справедливости нормального закона: экспериментаторы — потому, что они думают, что это математическая теорема; математики — потому, что они думают, что это экспериментальный факт».

Однако не следует упускать из виду, что полнота теоретических исследований, относящихся к нормальному закону, а также сравнительно простые математические свойства делают его наиболее привлекательным и удобным в применении. Даже в случае отклонения исследуемых экспериментальных данных от нормального закона существуют по крайней мере два пути его целесообразной эксплуатации: а) использовать его в качестве первого приближения; при этом нередко оказывается, что подобное допущение дает достаточно точные с точки зрения конкретных целей исследования результаты; б) подобрать такое преобразование исследуемой случайной величины ξ , которое видоизменяет исходный «ненормальный» закон распределения, превращая его в нормальный. Удобным для статистических приложений является и свойство «самовоспроизводимости» нормального закона, заключающееся в том, что сумма любого числа нормально распределенных случайных величин тоже подчиняется нормальному закону распределения. Кроме того, закон нормального распределения имеет большое теоретическое значение: с его помощью выведен целый ряд других важных распределений, построены различные статистические критерии и т. п. (χ^2 -, t - и F -распределения и опирающиеся на них критерии, см. п. 3.2.1–3.2.3, а также гл. 8).

Графики нормальных плотностей приведены на рис. 2.5, 2.6, 2.8 и 2.9.

Основные числовые характеристики нормального закона:

среднее, мода, медиана: $E\xi = x_{\text{mod}} = x_{\text{med}} = a;$

дисперсия: $D\xi = \sigma^2;$

$$\begin{array}{ll}
\text{асимметрия:} & \beta_1 = 0; \\
\text{эксцесс:} & \beta_2 = 0; \\
\text{центральные} & \\
\text{моменты} & m_k^{(0)} = \begin{cases} 0 & \text{при } k = 2m - 1, \\ 1 \cdot 3 \cdot \dots \cdot (2m - 1) \sigma^{2m} & \text{при } k = 2m \\ & m = 3, 4, \dots \end{cases} \\
\text{порядка } k \geq 3: &
\end{array}$$

Двумерный нормальный закон описывает совместное распределение двумерной случайной величины $\xi = (\xi^{(1)}, \xi^{(2)})$ с непрерывными компонентами $\xi^{(1)}$ и $\xi^{(2)}$, механизм формирования значений которых тот же, что и в одномерном случае, причем множества случайных факторов, под воздействием которых формируются значения $\xi^{(1)}$ и $\xi^{(2)}$, вообще говоря, пересекаются (отсюда возможная зависимость $\xi^{(1)}$ и $\xi^{(2)}$).

Введем в рассмотрение основные числовые характеристики двумерной случайной величины $\xi = (\xi^{(1)}, \xi^{(2)})$:

$$\text{вектор средних:} \quad M = \begin{pmatrix} m^{(1)} \\ m^{(2)} \end{pmatrix}, \text{ где } m^{(i)} = E\xi^{(i)};$$

$$\text{ковариационная матрица:} \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix},$$

$$\text{где } \sigma_{jk} = E\{(\xi^{(j)} - m^{(j)})(\xi^{(k)} - m^{(k)})\};$$

$$\text{коэффициент корреляции:} \quad r = \frac{\sigma_{12}}{(\sigma_{11}\sigma_{22})^{1/2}}.$$

Совместная двумерная плотность $\varphi(x^{(1)}, x^{(2)}; M, \Sigma)$ нормального закона может быть записана в виде

$$\begin{aligned}
\varphi(x^{(1)}, x^{(2)}; M, \Sigma) &= \frac{1}{2\pi[\sigma_{11}\sigma_{22}(1-r^2)]^{1/2}} \\
&\times \exp \left\{ -\frac{1}{2(1-r^2)} \left[\frac{(x^{(1)} - m^{(1)})^2}{\sigma_{11}} - 2r \frac{x^{(1)} - m^{(1)}}{\sigma_{11}^{1/2}} \cdot \frac{x^{(2)} - m^{(2)}}{\sigma_{22}^{1/2}} \right. \right. \\
&\quad \left. \left. + \frac{(x^{(2)} - m^{(2)})^2}{\sigma_{22}} \right] \right\} \quad (3.9)
\end{aligned}$$

или в виде

$$\varphi(x^{(1)}, x^{(2)}; M, \Sigma) = \varphi(X; M, \Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} e^{-\frac{1}{2}(X-M)^T \Sigma^{-1}(X-M)}, \quad (3.9')$$

где $X = \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix}$, верхний индекс Т означает транспонирование матрицы или вектора, $|\Sigma| = \det(\Sigma)$ — определитель ковариационной матрицы, Σ^{-1} — матрица, обратная к ковариационной, а $\exp\{c\} = e^c$. Изображение поверхности плотности двумерного нормального закона приведено на рис. 2.6.

Частные плотности $\varphi_{\xi^{(1)}}(x^{(1)}; m^{(1)}, \sigma_{11})$ и $\varphi_{\xi^{(2)}}(x^{(2)}; m^{(2)}, \sigma_{11})$ могут быть получены из совместной по формуле (2.12):

$$\varphi_{\xi^{(1)}}(x^{(1)}; m^{(1)}, \sigma_{11}) = \frac{1}{\sqrt{2\pi\sigma_{11}}} e^{-\frac{(x^{(1)} - m^{(1)})^2}{2\sigma_{11}}};$$

$$\varphi_{\xi^{(2)}}(x^{(2)}; m^{(2)}, \sigma_{22}) = \frac{1}{\sqrt{2\pi\sigma_{22}}} e^{-\frac{(x^{(2)} - m^{(2)})^2}{2\sigma_{22}}}.$$

Эти формулы означают, что частные законы распределения компонент двумерного нормального закона сами являются одномерными нормальными законами с параметрами соответственно $(m^{(1)}, \sigma_{11})$ и $(m^{(2)}, \sigma_{22})$.

Условные плотности $\varphi_{\xi^{(1)}}(x^{(1)} | \xi^{(2)} = x^{(2)})$ и $\varphi_{\xi^{(2)}}(x^{(2)} | \xi^{(1)} = x^{(1)})$ получаются с использованием общих формул (2.13) и (2.13'):

$$\begin{aligned} \varphi_{\xi^{(1)}}(x^{(1)} | \xi^{(2)} = x^{(2)}) &= \frac{1}{\sqrt{2\pi\sigma_{11}(1-r^2)}} \\ &\times e^{-\frac{1}{2\sigma_{11}(1-r^2)} \left[x^{(1)} - \left(m^{(1)} + r \frac{\sigma_{12}^{1/2}}{\sigma_{22}^{1/2}} (x^{(2)} - m^{(2)}) \right) \right]^2}; \end{aligned}$$

$$\begin{aligned} \varphi_{\xi^{(2)}}(x^{(2)} | \xi^{(1)} = x^{(1)}) &= \frac{1}{\sqrt{2\pi\sigma_{22}(1-r^2)}} \\ &\times e^{-\frac{1}{2\sigma_{22}(1-r^2)} \left[x^{(2)} - \left(m^{(2)} + r \frac{\sigma_{22}^{1/2}}{\sigma_{11}^{1/2}} (x^{(1)} - m^{(1)}) \right) \right]^2}. \end{aligned}$$

Отсюда следует, в частности, что условное распределение компоненты $\xi^{(i)}$ при фиксированном значении другой компоненты ($\xi^{(j)} = x^{(j)}$) снова описывается нормальным законом, параметр среднего значения которого, как и следовало ожидать, зависит от фиксированного значения $x^{(j)}$:

$$E(\xi^{(i)} | \xi^{(j)} = x^{(j)}) = m_i^{(i)} + r \frac{\sigma_{ii}^{1/2}}{\sigma_{jj}^{1/2}} (x^{(j)} - m_j^{(j)}),$$

и дисперсия которого не зависит от $x^{(j)}$ и равна

$$D(\xi^{(i)} | \xi^{(j)} = x^{(j)}) = \sigma_{ii}(1-r^2).$$

Многомерный нормальный закон описывает совместное распределение p -мерной случайной величины $\xi = (\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(p)})$ с непрерывными компонентами $\xi^{(j)}$, механизм формирования значений каждой из которых тот же, что и в одномерном случае, причем множества случайных факторов, под воздействием которых формируются значения $\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(p)}$, вообще говоря, пересекаются (отсюда их возможная взаимозависимость). Задавшись p -мерным вектор-столбцом M средних значений компонент и $(p \times p)$ -матрицей ковариаций Σ (см. п. 2.6.6), можно выпisać p -мерную совместную плотность многомерного нормального закона:

$$\begin{aligned} \varphi(x^{(1)}, x^{(2)}, \dots, x^{(p)}; M, \Sigma) \\ = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X-M)^T \Sigma^{-1}(X-M)}. \end{aligned} \quad (3.10)$$

Здесь, как и прежде, $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^T$ — вектор-столбец текущих переменных, а $|\Sigma| = \det(\Sigma)$ — определитель ковариационной матрицы.

Вырожденность матрицы Σ (т.е. равенство нулю определителя $|\Sigma|$) делает соответствующее многомерное распределение *вырожденным* (или *несобственным*); это означает, в частности, что разброс значений исследуемого многомерного признака сосредоточен в подпространстве меньшей, чем p , размерности. За исключением некоторых специальных случаев мы всегда будем полагать, что нами уже осуществлен переход в это подпространство меньшей размерности, так что в наших рассуждениях предполагается $|\Sigma| > 0$.

3.1.6. Логарифмически-нормальное распределение

Случайная величина η называется логарифмически-нормально распределенной, если ее логарифм $(\ln \eta)$ подчинен нормальному закону распределения. Это означает, в частности, что значения логарифмически-нормальной случайной величины формируются под воздействием очень большого числа взаимно независимых факторов, причем воздействие каждого отдельного фактора «равномерно незначительно» и равновероятно по знаку. При этом в отличие от схемы формирования механизма нормального закона последовательный характер воздействия случайных факторов таков, что случайный прирост, вызываемый действием каждого следующего фактора, пропорционален уже достигнутому к этому моменту значению исследуемой величины (в этом случае говорят о *мультипликативном* характере воздействия фактора). Математически сказанное может быть

формализовано следующим образом. Если $\eta_0 = a$ — *нелучайная* компонента исследуемого признака η (т. е. как бы «истинное» значение η в идеализированной схеме, когда устранено влияние всех случайных факторов), а $\xi_1, \xi_2, \dots, \xi_N$ — численное выражение эффектов воздействия упомянутых выше случайных факторов, то последовательно трансформированные действием этих факторов значения исследуемого признака будут:

$$\begin{aligned}\eta_1 &= \eta_0 + \xi_1 \cdot \eta_0; \\ \eta_2 &= \eta_1 + \xi_2 \cdot \eta_1; \\ &\dots\dots\dots \\ \eta_N &= \eta_{N-1} + \xi_N \cdot \eta_{N-1}.\end{aligned}$$

Отсюда легко получить

$$\sum_{i=0}^{N-1} \left(\frac{\Delta \eta_i}{\eta_i} \right) = \xi_1 + \xi_2 + \dots + \xi_N, \quad (3.11)$$

где $\Delta \eta_i = \eta_{i+1} - \eta_i$. Но правая часть (3.11) есть результат *аддитивного* действия множества случайных факторов, что при сделанных выше предположениях должно приводить, как мы знаем, к нормальному распределению этой суммы. В то же время, учитывая достаточную многочисленность числа случайных слагаемых (т.е. полагая $N \rightarrow \infty$) и относительную незначительность воздействия каждого из них (т.е. полагая $\Delta \eta_i \rightarrow 0$), можно от суммы в левой части (3.11) перейти к интегралу

$$\int_{\eta_0}^{\eta} \frac{d\eta}{\eta} = \ln \eta - \ln \eta_0 = \ln \eta - \ln a.$$

Это и означает в конечном счете, что логарифм интересующей нас величины (умноженный на постоянную величину $\ln a$) подчиняется нормальному закону с нулевым средним значением, т. е.

$$\begin{aligned}F_{\eta}(x) &= P\{\eta < x\} = P\{\ln \eta < \ln x\} \\ &= \frac{1}{\sqrt{2\pi} \cdot \sigma} \int_0^{\ln x} e^{-\frac{(t - \ln a)^2}{2\sigma^2}} dt,\end{aligned}$$

откуда дифференцированием по x левой и правой частей этого соотношения получаем

$$f_{\eta}(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma x} e^{-\frac{(\ln x - \ln a)^2}{2\sigma^2}} \quad (3.12)$$

(правомерность использованного при вычислении $f_\eta(x)$ тождества $P\{\eta < x\} = P\{\ln \eta < \ln x\}$ вытекает из строгой монотонности преобразования $\ln \eta$).

Описанная схема формирования значений логарифмически-нормальной случайной величины оказывается характерной для многих конкретных физических и социально-экономических ситуаций (размеры и вес частиц, образующихся при дроблении; заработная плата работника; доход семьи; размеры космических образований; долговечность изделия, работающего в режиме износа и старения, и др). Примеры графиков плотности логарифмически-нормального распределения представлены на рис. 3.4

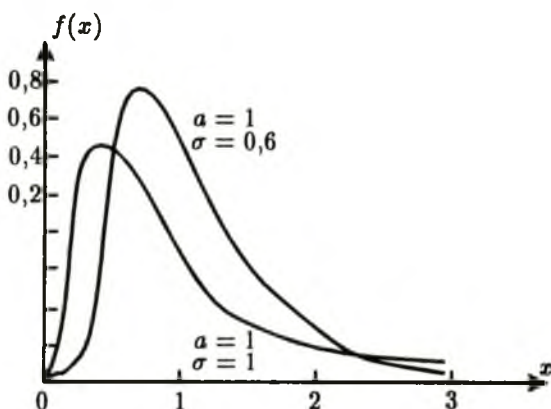


Рис. 3.4. График функции плотности логарифмически-нормального распределения для различных σ при $a = 1$

Ниже приводятся результаты вычисления основных числовых характеристик логарифмически-нормального распределения (в терминах параметров закона a и σ^2):

среднее:	$E\eta = ae^{\frac{1}{2}\sigma^2};$
мода:	$x_{\text{mod}} = ae^{-\sigma^2};$
медиана:	$x_{\text{med}} = a;$
дисперсия:	$D\eta = (E\eta)^2(e^{\sigma^2} - 1) = a^2 e^{\sigma^2}(e^{\sigma^2} - 1);$
асимметрия:	$\beta_1 = (e^{\sigma^2} - 1)^{\frac{1}{2}}(e^{\sigma^2} + 2);$
эксцесс:	$\beta_2 = (e^{\sigma^2} - 1)(e^{3\sigma^2} + 3e^{2\sigma^2} + 6e^{\sigma^2} + 6).$

Из этих выражений видно, что асимметрия и эксцесс логарифми-

чески-нормального распределения всегда положительны (и тем ближе к нулю, чем ближе к нулю σ^2), а мода, медиана и среднее выстраиваются как раз в том порядке, который мы видим на рис. 2.7, причем они будут стремиться к слиянию (а кривая плотности — к симметрии) по мере стремления к нулю величины σ^2 . При этом, хотя значения логарифмически-нормальной случайной величины образуются как «случайные искажения» некоторого «истинного значения» a , последнее в конечном счете выступает не в роли среднего значения, а в роли медианы.

3.1.7. Равномерное (прямоугольное) распределение

Случайная величина ξ называется равномерно распределенной на отрезке $[a, b]$, если ее плотность вероятности $f_\xi(x)$ постоянна на этом отрезке и равна нулю вне его, т. е.

$$f_\xi(x) = \begin{cases} \frac{1}{b-a} & \text{при } a \leq x \leq b; \\ 0 & \text{при } a < a \text{ и } x > b. \end{cases} \quad (3.13)$$

Так как график функции $f_\xi(x)$ изображается в виде прямоугольника (см. рис. 3.5), то такое распределение также называют *прямоугольным*.

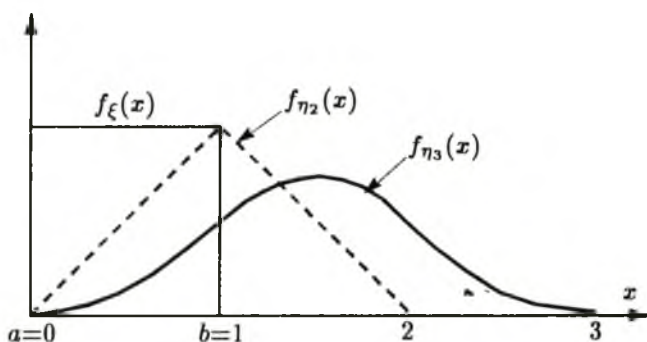


Рис. 3.5. Функция плотности равномерной случайной величины ($f_\xi(x)$) и суммы двух ($f_{\eta_2}(x)$) и трех ($f_{\eta_3}(x)$) независимых равномерно распределенных на $[0,1]$ случайных величин

Соответственно функция распределения $F_\xi(x)$ равномерного закона задается соотношениями:

$$F_{\xi}(x) = \begin{cases} 0 & \text{при } x \leq a; \\ \frac{x-a}{b-a} & \text{при } a < x \leq b; \\ 1 & \text{при } x > b. \end{cases} \quad (3.13')$$

Примерами реальных ситуаций, связанных с необходимостью рассмотрения равномерно распределенных случайных величин, могут служить: анализ ошибок округления при проведении числовых расчетов (такая ошибка, как правило, оказывается равномерно распределенной на интервале от -5 до $+5$ единиц округляемого десятичного знака); время ожидания «обслуживания» при точно периодическом, через каждые T единиц времени, включении (прибытии) «обслуживающего устройства» и при случайном поступлении (прибытии) заявки на обслуживание в этом интервале (например, время ожидания пассажира прибытия поезда метро при условии точных двухминутных интервалов движения и случайного момента появления пассажира на платформе будет распределено приблизительно равномерно на интервале $[0 \text{ мин}, 2 \text{ мин}]$).

Отметим еще две важные ситуации, в которых используется равномерный закон. Во-первых, в теории и практике статистического анализа данных широко используется вспомогательный переход от исследуемой случайной величины ξ с функцией распределения $F(x)$ к случайной величине $\eta = F(\xi)$, которая оказывается равномерно распределенной на отрезке $[0,1]$ (см. п. 4.4). Этот прием является полезным при *статистическом моделировании наблюдений*, подчиненных заданному закону распределения вероятностей, при построении доверительных границ для исследуемой функции распределения и в ряде других задач математической статистики. Во-вторых, равномерное распределение иногда используется в качестве «нулевого приближения» в описании априорного распределения анализируемых параметров в так называемом *байесовском подходе* в условиях полного отсутствия априорной информации об этом распределении (см. п. 7.6.).

Числовые характеристики равномерного закона:

$$\text{среднее, медиана:} \quad E\xi = x_{\text{med}} = \frac{a+b}{2};$$

$$\text{дисперсия:} \quad D\xi = \frac{(b-a)^2}{12};$$

$$\text{асимметрия:} \quad \beta_1 = 0;$$

$$\text{эксцесс:} \quad \beta_2 = -1,2.$$

Отметим в заключение одно важное свойство суммы n независимых равномерно распределенных случайных величин: распределение этой суммы *очень быстро* (по мере роста числа слагаемых) приближается к нормальному закону. В частности, если ξ_i — равномерно распределенные на отрезке $[0,1]$ и независимые случайные величины, то можно показать, что плотность $f_{\eta_n}(x)$ случайной величины $\eta_n = \xi_1 + \dots + \xi_n$ имеет вид

$$f_{\eta_n}(x) = \begin{cases} \frac{1}{(n-1)!} x^{n-1} & \text{при } 0 \leq x \leq 1; \\ \frac{1}{(n-1)!} [x^{n-1} - C_n^1(x-1)^{n-1}] & \text{при } 1 < x \leq 2; \\ \frac{1}{(n-1)!} [x^{n-1} - C_n^1(x-1)^{n-1} + C_n^2(x-2)^{n-1}] & \text{при } 2 < x \leq 3; \\ \dots & \dots \\ \frac{1}{(n-1)!} [x^{n-1} - C_n^1(x-1)^{n-1} + \dots + (-1)^{n-1} C_n^{n-1}(x-(n-1))^{n-1}] & \text{при } n-1 < x \leq n. \end{cases}$$

(область возможных значений случайной величины η_n , очевидно, задается отрезком $[0, n]$). Геометрическое изображение последовательного изменения вида плотности $f_{\eta_n}(x)$ по мере роста числа слагаемых n (для $n = 1, 2, 3$) дано на рис. 3.5. Это свойство используется, в частности, при статистическом моделировании нормально распределенных наблюдений.

3.1.8. Распределения Вейбулла и экспоненциальное (показательное)

Рассмотрим один общий механизм формирования распределений, относящийся, в частности, к случайным величинам, характеризующим длительность жизни элемента, сложной системы или индивидуума (задачи теории надежности, анализ коэффициентов смертности в демографии и т. п.). Пусть ξ — время жизни анализируемого объекта (системы, индивидуума) и $F(t) = P\{\xi < t\}$ — его функция распределения, которую мы полагаем непрерывной и дифференцируемой. В задачах данного профиля важной характеристикой является интенсивность отказов (*коэффициент смертности*) $\lambda(t)$ исследуемых элементов возраста t , определяемая соотношением

$$\lambda(t) = \frac{f_{\xi}(t)}{1 - F_{\xi}(t)}. \quad (3.14)$$

Статистически (экспериментально) интенсивность отказов определяется как отношение удельного числа (т. е. приходящегося на единицу времени) выбывших в возрасте t элементов к общему числу доживших до

этого возраста элементов $n(t)$, т. е.:

$$\hat{\lambda}(t) = \frac{n(t) - n(t + \Delta t)}{\Delta t \cdot n(t)} = \frac{\frac{[n - n(t + \Delta t)]}{\Delta t \cdot n} - \frac{[n - n(t)]}{\Delta t \cdot n}}{\frac{n - (n - n(t))}{n}} = \frac{\hat{f}_{\xi}(t)}{1 - \hat{F}_{\xi}(t)},$$

где n — общее число наблюдаемых во времени (начиная с $t = 0$) элементов одинакового возраста, а $\hat{f}_{\xi}(t)$ и $\hat{F}_{\xi}(t)$, соответственно, эмпирическая плотность и эмпирическая функция распределения анализируемой случайной величины ξ .

Разрешая уравнение (3.14) относительно функции распределения $F_{\xi}(t)$, получаем

$$F_{\xi}(x) = 1 - e^{-\int_0^x \lambda(t) dt}. \quad (3.15)$$

Таким образом, конкретизация вида функции распределения $F_{\xi}(x)$ полностью определяется видом функции интенсивности отказов (временной зависимостью коэффициентов смертности) $\lambda(t)$.

Многочисленные экспериментальные данные (и в области демографии, и в области анализа надежности технических элементов и систем) показывают, что в широком классе случаев функция $\lambda(t)$ имеет характерный вид кривой, изображенной на рис. 3.6. Из этого графика видно, что весь интервал времени можно разбить на три периода.



Рис. 3.6. Типичное поведение кривой смертности (интенсивности отказов): кривая описывает изменение коэффициента смертности мужского населения Франции по данным 1955 г.¹

На первом из них функция $\lambda(t)$ имеет высокие значения и явную тенденцию к убыванию. На техническом языке это можно объяснить наличием в исследуемой совокупности элементов с явными и скрытыми дефектами (сборки, некондиционности отдельных свойств и т. п.), которые приводят к относительно быстрому выходу из строя этих элементов. Этот период принято называть *периодом «приработки»* (или «*обкатки*»). Затем наступает период нормальной эксплуатации, характеризующийся приблизительно постоянным и сравнительно низким уровнем «смертности» элементов. Природа смертей (или «отказов») в этот период носит внезапный характер (аварии, несчастные случаи и т. п.) и не зависит от возраста объекта. И наконец, последний период жизни (или эксплуатации) элемента — *период старения и износа*. Природа «отказов» в этот период — в необратимых физиологических или физико-химических явлениях, приводящих к ухудшению качества элемента, к его «старению». В соответствии с этой кривой смертности периоду «приработки» соответствует возраст от 0 до 9 лет, периоду «нормальной эксплуатации» — от 9 до 39 лет, а периоду «старения» — возраст после 39 лет.

Каждому периоду соответствует свой вид функции $\lambda(t)$, а следовательно, и свой закон распределения времени жизни ξ .

Рассмотрим класс *степенных зависимостей*, описывающих поведение $\lambda(t)$, а именно

$$\lambda(t) = \lambda_0 \alpha t^{\alpha-1}, \quad (3.16)$$

где $\lambda_0 > 0$ и $\alpha > 0$ — некоторые числовые параметры. Очевидно, значения $\alpha < 1$, $\alpha = 1$ и $\alpha > 1$ отвечают поведению функции интенсивности отказов соответственно в период приработки, нормальной эксплуатации и старения.

Подставляя (3.16) в (3.15), получаем вид функции распределения $F_\xi(t)$ времени жизни ξ элемента:

$$F_\xi(t) = 1 - e^{-\lambda_0 t^\alpha}, \quad t \geq 0. \quad (3.17)$$

Соответственно плотность вероятности

$$f_\xi(t) = \lambda_0 \alpha t^{\alpha-1} e^{-\lambda_0 t^\alpha}, \quad t \geq 0. \quad (3.17')$$

Это и есть *распределение Вейбулла*. Поведение графиков функции плотности распределения Вейбулла представлено на рис.3.7. К тому же самому типу распределения можно прийти, отправляясь от широкого класса различных вероятностных законов и интересуясь распределением крайних членов их вариационных рядов.

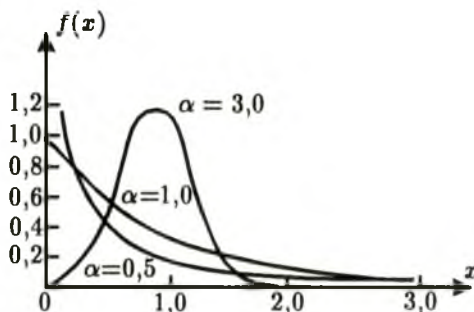


Рис. 3.7. Графики функций плотности распределения Вейбулла для различных значений α

Основные числовые характеристики распределения Вейбулла:

среднее:
$$E\xi = \lambda_0^{-\frac{1}{\alpha}} \Gamma\left(1 + \frac{1}{\alpha}\right);$$

мода:
$$x_{\text{mod}} = \begin{cases} 0, & \text{если } \alpha \leq 1; \\ \lambda_0^{-\frac{1}{\alpha}} \left(1 - \frac{1}{\alpha}\right)^{\frac{1}{\alpha}}, & \text{если } \alpha \geq 1; \end{cases}$$

дисперсия:
$$D\xi = \lambda_0^{-\frac{2}{\alpha}} \cdot \left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right) \right];$$

момент k -го порядка:
$$m_k = E\xi^k = \lambda_0^{-\frac{k}{\alpha}} \cdot \Gamma\left(1 + \frac{k}{\alpha}\right)$$

(здесь $\Gamma(z)$ — так называемая гамма-функция Эйлера, т.е. $\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx$).

Экспоненциальное (показательное) распределение хотя и является частным случаем распределения Вейбулла (когда $\alpha = 1$, см. соответствующую кривую плотности на рис. 3.7), но представляет большой самостоятельный интерес. Из вышесказанного следует, что оно адекватно описывает распределение длительности жизни элемента, работающего в режиме нормальной эксплуатации. Экспоненциальный закон (и только он) обладает, в частности, тем важным свойством, что вероятность безотказной работы элемента на данном временном интервале $(t, t + \Delta)$ не зависит от времени предшествующей работы t , а зависит только от длины интервала Δ . Экспоненциально распределенную случайную величину можно интерпретировать так же, как промежуток времени между двумя последовательными наступлениями «пуассоновского» события. Прикладная популярность экспоненциального закона объясняется не только разно-

образными возможностями его естественной физической интерпретации, но и исключительной простотой и удобством его модельных свойств. Ниже приводятся его функция распределения и плотность вероятности, а также его основные числовые характеристики:

$$F_{\xi}(x) = 1 - e^{-\lambda_0 x}, \quad x \geq 0;$$

$$f_{\xi}(x) = \lambda_0 e^{-\lambda_0 x}, \quad x \geq 0;$$

<i>среднее:</i>	$E\xi = \frac{1}{\lambda_0};$
<i>мода:</i>	$x_{\text{mod}} = 0;$
<i>медиана:</i>	$x_{\text{med}} = \frac{1}{\lambda_0} \cdot \ln 2;$
<i>дисперсия:</i>	$D\xi = \frac{1}{\lambda_0^2};$
<i>асимметрия:</i>	$\beta_1 = 2;$
<i>эксцесс:</i>	$\beta_2 = 6.$

Двустороннее экспоненциальное распределение (распределение Лапласа). Симметричная унимодальная функция плотности этого закона с «острым» максимумом в точке $x = 0$ часто используется для описания распределения остаточных случайных компонент («ошибок») ε в моделях регрессионного типа (см. гл. 11.). График этой функции плотности представляет собой как бы результат «склеивания» графика показательного распределения со своим зеркальным — относительно вертикальной оси — отражением (с учетом необходимой нормировки), так что уравнение функции плотности имеет вид

$$f(x) = \frac{1}{2} \lambda \cdot e^{-\lambda \cdot |x|} \quad (-\infty < x < \infty).$$

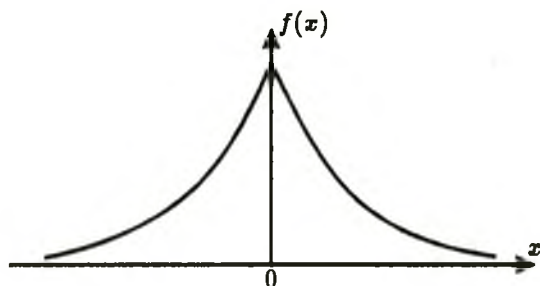


Рис. 3.8. График функции плотности распределения Лапласа

Нетрудно подсчитать основные числовые характеристики этого закона распределения:

<i>среднее:</i>	$E\xi = 0;$
<i>мода:</i>	$x_{\text{mod}} = 0;$
<i>медиана:</i>	$x_{\text{med}} = 0;$
<i>дисперсия:</i>	$D\xi = 2/\lambda^2;$
<i>асимметрия:</i>	$\beta_1 = 0;$
<i>эксцесс:</i>	$\beta_2 = 3.$

3.1.9. Распределение Парето

В различных задачах прикладной статистики довольно часто встречаются так называемые «усеченные» распределения. Такие распределения описывают вероятностные закономерности в неполных («усеченных») генеральных совокупностях, т. е. в таких совокупностях, из которых заранее изъяты все элементы с количественным признаком, превышающим некоторый заданный уровень c_0 (или, наоборот, меньшим, чем c_0). Скажем, налоговые органы обычно интересуются распределением годовых доходов тех лиц, годовой доход которых превосходит некоторый порог c_0 , установленный законами о налогообложении. Эти и некоторые аналогичные распределения иногда оказываются приблизительно совпадающими с распределением Парето, задаваемым функциями:

$$F_{\xi}(x) = P\{\xi < x\} = 1 - \left(\frac{c_0}{x}\right)^{\alpha},$$

$$f_{\xi}(x) = \frac{\alpha}{c_0} \left(\frac{c_0}{x}\right)^{\alpha+1}.$$

В этих формулах $\alpha > 0$, а $x > c_0$, т. е. область мыслимых значений данной случайной величины ξ есть полупрямая $(c_0, +\infty)$. Функция плотности имеет вид монотонно убывающей кривой, выходящей из точки $(c_0, \alpha/c_0)$.

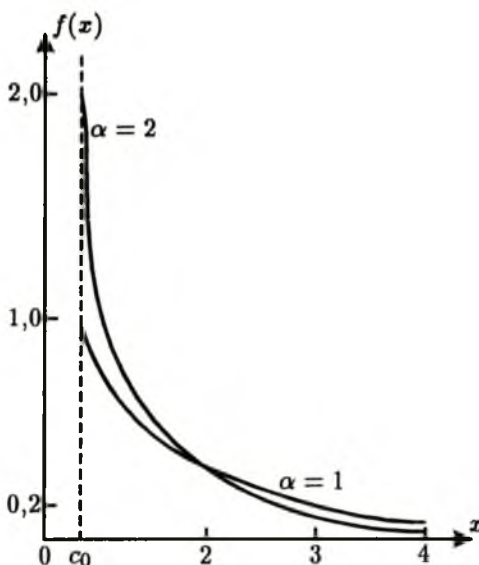


Рис. 3.9. График функции плотности распределения Парето для различных значений α

Основные числовые характеристики этого распределения существуют не всегда, а лишь при соблюдении определенных требований к величине параметра α :

среднее: $E = \frac{\alpha}{\alpha - 1} c_0$ (существует при $\alpha > 1$);

мода: $x_{\text{mod}} = c_0$;

медиана: $x_{\text{med}} = 2^{\frac{1}{\alpha}} \cdot c_0$;

дисперсия: $D\xi = \frac{\alpha}{(\alpha - 1)^2(\alpha - 2)} c_0^2$ (существует при $\alpha > 2$);

момент k -го порядка: $E\xi^k = \frac{\alpha}{\alpha - k} c_0^k$ (существует при $\alpha > k$).

3.1.10. Распределение Коши

Этот закон весьма специфичен, поскольку ни один из его моментов положительного порядка (в том числе даже среднее значение) не суще-

ствуется. Распределение Коши унимодально, симметрично относительно своего модального значения (которое, следовательно, одновременно является медианой) и задается функцией плотности вероятности

$$f(x) = \frac{1}{\pi} \cdot \frac{c}{c^2 + (x - a)^2}, -\infty < x < \infty,$$

где $c > 0$ — параметр масштаба и a — параметр центра группирования, определяющий одновременно значение моды и медианы.

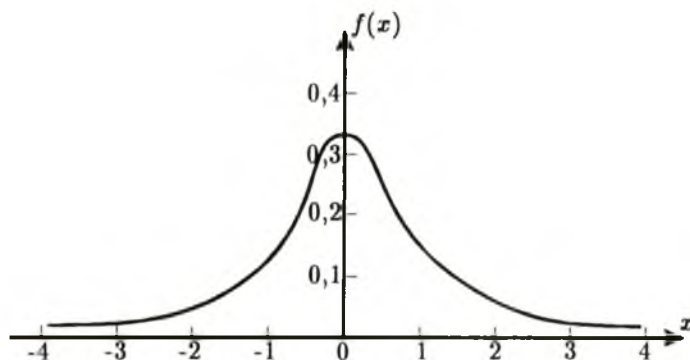


Рис. 3.10. График функции плотности распределения Коши

Соответственно функция распределения задается соотношением

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \operatorname{arcth} \frac{x - a}{c}.$$

Отметим два важных свойства («самовоспроизводимости») распределения Коши.

1. Если случайная величина ξ имеет распределение Коши с параметрами c и a , то любая линейная функция $b_0 + b_1\xi$ имеет распределение того же типа с параметрами $c' = |b_1| \cdot c$ и $a' = b_1 \cdot a + b_0$;

2. Если случайные величины $\xi_1, \xi_2, \dots, \xi_n$ независимы и имеют одно и то же распределение Коши, то среднее арифметическое $\bar{\xi} = (\xi_1 + \dots + \xi_n)/n$ имеет то же самое распределение, что и каждое ξ_i .

3.2. Законы распределения вероятностей, используемые при реализации техники статистических вычислений

Ниже описываются пять законов, основное назначение которых — предоставление исследователю необходимого аппарата для построения

разного рода статистических критериев и интервальных оценок параметров: χ^2 -распределение («хи-квадрат»-распределение); t -распределение (Стьюдента); F -распределение (распределение дисперсионного отношения); B -распределение (бета-распределение) и Γ -распределение (гамма-распределение). Объяснение механизма их действия можно связывать со «статистикой нормального закона», т. е. с изучением распределений некоторых функций от набора независимых и одинаково стандартно нормально распределенных случайных величин.

3.2.1. «Хи-квадрат»-распределение с m степенями свободы ($\chi^2(m)$ -распределение)

В связи с гауссовской теорией ошибок астроном Ф. Хельмерт исследовал суммы квадратов m независимых стандартно нормально распределенных случайных величин, придя таким образом к функции распределения $F_{\chi^2(m)}(x)$, которую позднее К. Пирсон назвал функцией распределения «хи-квадрат»¹. Для отрицательных x функция $F_{\chi^2(m)}(x) = 0$, а для неотрицательных x

$$\begin{aligned} F_{\chi^2(m)}(x) &= P\{\chi^2(m) < x\} \\ &= \frac{1}{2^{\frac{m}{2}} \Gamma(\frac{m}{2})} \cdot \int_0^x t^{\frac{m}{2}-1} e^{-\frac{t}{2}} dt, \end{aligned} \quad (3.18)$$

где параметр закона m — целое положительное число, которое принято называть *числом степеней свободы*, а $\Gamma(y)$ — значение гамма-функции Эйлера в точке y .²

Соответствующая плотность вероятности задается функцией

$$f_{\chi^2(m)}(x) = \frac{1}{2^{\frac{m}{2}} \Gamma(\frac{m}{2})} \cdot x^{\frac{m}{2}-1} e^{-\frac{x}{2}}, \quad x > 0. \quad (3.18')$$

¹ Helmer F. R. Über die Wahrscheinlichkeit von Potenzsummen der Beobachtungsfehler etc. Z. F. Math. und Phys., 21 (1876); Pearson K. On the criterion that a given system of deviation from the probable in the case of correlated system of variables..., «Phil Mag.», V. 50 (1900), 157.

² Значение гамма-функции Эйлера в точке y определяется интегралом: $\Gamma(y) = \int_0^\infty e^{-t} t^{y-1} dt$. Из определения непосредственно следует, например, что при целых положительных значениях y : $\Gamma(y) = (y-1)!$ и $\Gamma(y + \frac{1}{2}) = 2^{-y} \sqrt{\pi} \cdot 1 \cdot 3 \cdot 5 \cdots (2y-1)$ (более подробные сведения о гамма-функции см. в книге [Большев Л. Н., Смирнов Н. В.]).

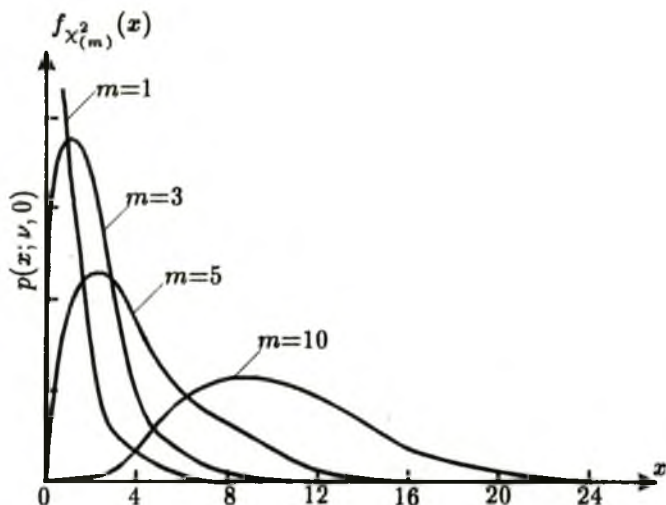


Рис. 3.11. График функции плотности $\chi^2(m)$ -распределения для различных значений m

Распределение χ^2 появилось впервые при исследовании распределения последовательности независимых и одинаково стандартно нормально распределенных случайных величин $\xi_1, \xi_2, \dots, \xi_m$. Выяснилось, что случайная величина $\chi^2(m) = \xi_1^2 + \dots + \xi_m^2$ подчиняется закону χ^2 -распределения с m степенями свободы.

Основные числовые характеристики $\chi^2(m)$ -распределения:

среднее:	$E\chi^2(m) = m;$
мода:	$m_{\text{mod}} = m - 2 \quad (m \geq 2);$
дисперсия:	$D\chi^2(m) = 2m;$
асимметрия:	$\beta_1 = \frac{2^{3/2}}{\sqrt{m}};$
эксцесс:	$\beta_2 = \frac{12}{m}.$

3.2.2. Распределение Стьюдента с m степенями свободы ($t(m)$ -распределение)

Английский статистик В. Госсет (писавший под псевдонимом «Стью-

дент») получил в 1908 г.¹ следующий результат. Пусть $\xi_0, \xi_1, \dots, \xi_m$ — независимые $(0, \sigma^2)$ -нормально распределенные случайные величины. Тогда плотность распределения случайной величины

$$t(m) = \frac{\xi_0}{\left(\frac{1}{m} \sum_{i=1}^m \xi_i^2 \right)^{\frac{1}{2}}}$$

описывается функцией

$$f_t(x) = \frac{1}{\sqrt{\pi m}} \cdot \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \cdot \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}} \quad (3.19)$$

$(-\infty < x < +\infty).$

Распределение (3.19) получило название *распределения Стьюдента с m степенями свободы* (или $t(m)$ -распределения). Из выражения (3.19) следует, что функция плотности $f_t(x)$ не зависит от дисперсии σ^2 случайных величин ξ_i и, кроме того, является унимодальной и симметричной относительно точки $x = 0$, (см. рис. 3.12).

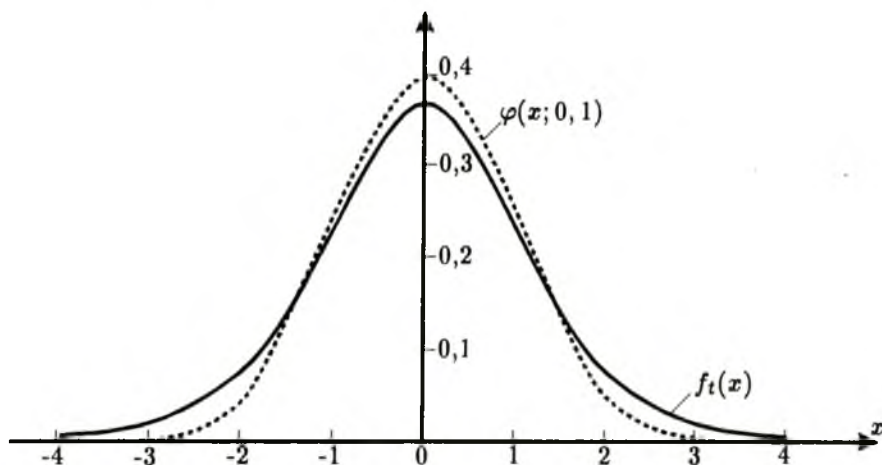


Рис. 3.12. Графики функции плотности $t(m)$ -распределения при $m = 4$ и стандартного нормального распределения $\varepsilon(x; 0, 1)$

Основные числовые характеристики $t(m)$ -распределения:

среднее, мода, медиана: $Et(m) = x_{\text{mod}} = x_{\text{med}} = 0$;

¹ S t u d e n t . The probable error of a mean. — Biometrika, B, 6 (1908), 1.

дисперсия:	$Dt(m) = \frac{m}{m-2}$ (существует только при $m > 2$);
асимметрия:	$\beta_1 = 0$;
эксцесс:	$\beta_2 = \frac{6}{m-4}$ (существует только при $m > 4$).

3.2.3. Распределение дисперсионного отношения с числом степеней свободы числителя m_1 и числом степеней свободы знаменателя m_2 ($F(m_1, m_2)$ -распределение)

Рассмотрим $m_1 + m_2$ независимых и $(0, \sigma^2)$ -нормально распределенных величин $\xi_1, \dots, \xi_{m_1}$; $\eta_1, \dots, \eta_{m_2}$ и положим

$$F(m_1, m_2) = \frac{\frac{1}{m_1} \sum_{i=1}^{m_1} \xi_i^2}{\frac{1}{m_2} \sum_{j=1}^{m_2} \eta_j^2}.$$

Очевидно (см. п. 3.2.1), та же самая случайная величина может быть определена и как отношение двух *независимых* и соответствующим образом нормированных χ^2 -распределенных величин $\chi^2(m_1)$ и $\chi^2(m_2)$, т. е.

$$F(m_1, m_2) = \frac{\frac{1}{m_1} \chi^2(m_1)}{\frac{1}{m_2} \chi^2(m_2)}.$$

Английский статистик Р. Фишер в 1924 году показал¹, что плотность вероятности случайной величины $F(m_1, m_2)$ задается функцией

$$f_{F(m_1, m_2)}(x) = \frac{\Gamma\left(\frac{m_1 + m_2}{2}\right) \cdot m_1^{\frac{m_1}{2}} \cdot m_2^{\frac{m_2}{2}}}{\Gamma\left(\frac{m_1}{2}\right) \cdot \Gamma\left(\frac{m_2}{2}\right)} \times \frac{x^{\frac{m_1}{2} - 1}}{(m_1 x + m_2)^{\frac{m_1 + m_2}{2}}} \quad (0 \leq x < \infty), \quad (3.20)$$

¹ Fisher R. On a distribution yielding the error functions of several well-known statistics. — Proc. Intern. Math. Congr. Toronto, 1924, 805.

где, как обычно, $\Gamma(y)$ — значение гамма-функции Эйлера в точке y , а сам закон называется F -распределением с числами степеней свободы числителя и знаменателя, равными соответственно m_1 и m_2 .

Основные числовые характеристики $F(m_1, m_2)$ -распределения:

среднее: $EF(m_1, m_2) = \frac{m_2}{m_2 - 2}$ (существует при $m_2 > 2$);

мода: $x_{\text{mod}} = \frac{(m_1 - 2) \cdot m_2}{m_1(m_2 + 2)}$ (для $m_1 > 1$);

дисперсия: $DF(m_1, m_2) = \frac{2m_2^2(m_1 + m_2 - 2)}{m_1(m_2 - 2)^2(m_2 - 4)}$ (при $m_2 > 4$);

асимметрия: $\beta_1 = \frac{(2m_1 + m_2 - 2) \sqrt{8(m_2 - 4)}}{(m_2 - 6) \sqrt{(m_1 + m_2 - 2)m_1}}$ (при $m_2 > 6$);

эксцесс: $\beta_2 = \frac{3(m_2 - 6) \left(2 + \frac{1}{2}\beta_1^2\right)}{m_2 - 8} - 3$ (при $m_2 > 8$).

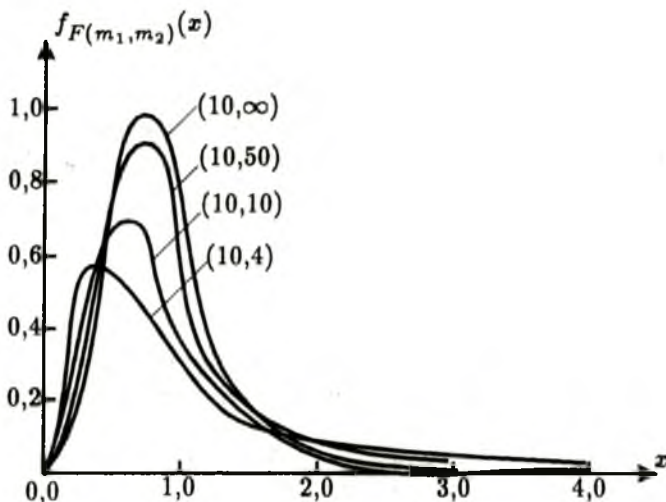


Рис. 3.13. График функции плотности $F(m_1, m_2)$ -распределения для различных значений m_2 при $m_1 = 10$

Отсюда непосредственно следует, что при $m_1, m_2 > 2$ F -распределение всегда имеет модальное значение, меньшее единицы, и среднее значение, большее единицы. Это означает, в частности, что данное распределение имеет асимметрию, подобную той, что изображена на рис. 2.6 (в

этих случаях говорят, что распределение имеет *положительную асимметрию* независимо от того, существует ли коэффициент асимметрии β_1).

Р. Фишером данный тип распределений выводился не для случайной величины $F(m_1, m_2)$, а для ее натурального логарифма (поделенного пополам), т. е. для случайной величины $\zeta = \frac{1}{2} \ln F(m_1, m_2)$. Распределение этой случайной величины часто называют *z-распределением Фишера*. Однако в современной статистической практике предпочитают использовать *F-распределение*, обладающее более простыми свойствами.

3.2.4. Гамма-распределение (Г-распределение)

Описанные ниже два типа распределения представляют весьма широкие и гибкие двухпараметрические семейства законов, которые включают в себя в качестве частных случаев различные комбинации уже известных нам случайных величин. Основное прикладное значение Г- и В-распределений — *в их богатых вычислительных возможностях*: функции распределения χ^2 , t , F и ряд других могут быть вычислены (после подходящего преобразования переменных) в терминах Г- или В-распределений. Кроме того, Г-распределение иногда используется и при реалистическом моделировании: с его помощью описывается, например, распределение доходов и сбережений населения в определенных специальных ситуациях.

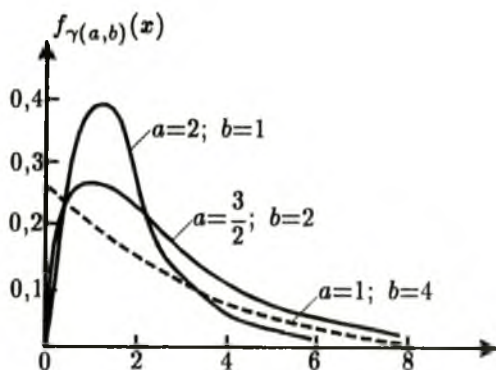


Рис. 3.14. График функции плотности гамма-распределения для различных значений a и b

Двухпараметрический закон Г-распределения случайной величины

$\gamma(a, b)$ описывается функцией плотности:

$$f_{\gamma(a,b)}(x) = \begin{cases} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} & \text{при } 0 \leq x < \infty; \\ 0 & \text{при } x < 0 \end{cases} \quad (3.21)$$

где $\Gamma(a)$ — Γ -функция Эйлера, $a > 0$ — параметр «формы» и $b > 0$ — параметр масштаба. Легко понять, что Γ -распределенная случайная величина с параметрами a и b (будем обозначать ее $\gamma(a, b)$) связана со случайной величиной $\gamma(a, 1)$ простым соотношением:

$$b\gamma(a, b) = \gamma(a, 1).$$

Отметим несколько полезных свойств Γ -распределения.

1. Из вида функции плотности $\chi^2(m)$ -распределения (см. (3.18)) непосредственно следует, что оно является частным случаем Γ -распределения: достаточно положить в (3.21) $a = m/2$ и $b = 1/2$.

2. Сумма любого числа независимых Γ -распределенных случайных величин (с одинаковым параметром масштаба b) $\gamma_1(a_1, b) + \gamma_2(a_2, b) + \dots + \gamma_n(a_n, b)$ также подчиняется Γ -распределению, но с параметрами $a_1 + a_2 + \dots + a_n$ и b .

Основные числовые характеристики случайной величины $\gamma(a, b)$:

<i>среднее:</i>	$E\gamma(a, b) = \frac{a}{b};$
<i>мода:</i>	$x_{\text{mod}} = \frac{a-1}{b} \quad (\text{при } a \geq 1);$
<i>дисперсия:</i>	$D\gamma(a, b) = \frac{a}{b^2};$
<i>асимметрия:</i>	$\beta_1 = \frac{2}{\sqrt{a}};$
<i>эксцесс:</i>	$\beta_2 = \frac{6}{a}.$

3.2.5. Бета-распределение (B -распределение)

Как отмечалось выше, двухпараметрический закон B -распределения обладает весьма высокой гибкостью и общностью: в частности, через функцию B -распределения могут быть вычислены такие часто используемые распределения, как t^2 , F , биномиальное, отрицательное биномиальное и др. B -распределение используется и для описания некоторых реальных

распределений, сосредоточенных на отрезке $[0,1]$ (например, для описания распределения величин субъективных вероятностей, полученных в ходе экспертного опроса, см. п. 2.1.3). Случайная величина $\beta(a_1, a_2)$, подчиняющаяся закону B -распределения с параметрами a_1 и a_2 ($0 < a_1 < \infty$, $0 < a_2 < \infty$), имеет плотность вероятности

$$f_{\beta(a_1, a_2)}(x) = \begin{cases} \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \cdot x^{a_1-1}(1-x)^{a_2-1} & \text{при } 0 \leq x \leq 1; \\ 0 & \text{для остальных значений } x. \end{cases} \quad (3.22)$$

Отметим несколько полезных свойств B -распределения.

1. Если $\gamma(a_1, b)$ и $\gamma(a_2, b)$ — две независимые Γ -распределенные случайные величины, то отношение $\beta(a_1, a_2) = \gamma(a_1, b)/(\gamma(a_1, b) + \gamma(a_2, b))$ подчиняется закону B -распределения с параметрами a_1 и a_2 .

2. Случайная величина $\beta(1, 1)$ распределена равномерно на отрезке $[0,1]$ (см. п. 3.1.7).

3. Функция распределения квадрата стьюдентовской величины $t^2(m)$ (см. п. 3.2.2) связана с функцией распределения случайной величины β соотношением

$$F_{t^2(m)}(x) = F_{\beta(\frac{1}{2}, \frac{m}{2})} \left(\frac{x^2}{m + x^2} \right).$$

4. Функция распределения случайной величины $F(m_1, m_2)$ (см. п. 3.2.3) связана с функцией распределения случайной величины β соотношением

$$F_{F(m_1, m_2)}(x) = F_{\beta(\frac{m_1}{2}, \frac{m_2}{2})} \left(\frac{m_1 x}{m_2 + m_1 x} \right).$$

5. Между функцией распределения случайной величины β и распределениями биномиального типа (см. п. 3.1.1) существуют соотношения:

$$F_{\beta(m, n-m+1)}(x) = \sum_{k=m}^n C_n^k x^k (1-x)^{n-k};$$

$$F_{\beta(n, m)}(1-x) = \sum_{k=n}^{\infty} C_{m+k-1}^{m-1} x^m (1-x)^k.$$

6. Непосредственный анализ плотности (3.22) обнаруживает симметричность плотностей $f_{\beta(a_1, a_2)}(x)$ и $f_{\beta(a_2, a_1)}(x)$ относительно прямой $x = 0,5$ (рис. 3.15), что в терминах соответствующих функций распределения

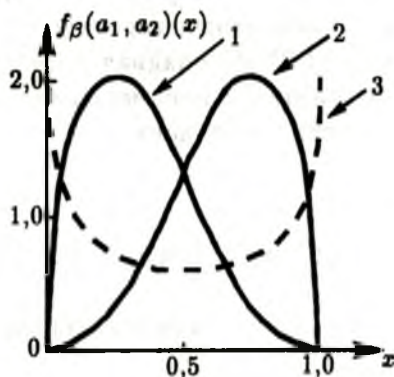


Рис. 3.15. Графики функций плотности β -распределения при различных значениях параметров a_1 и a_2 : 1) $a_1 = 2$, $a_2 = 4$; 2) $a_1 = 4$, $a_2 = 2$; 3) $a_1 = 1/2$, $a_2 = 1/2$

может быть записано в виде

$$F_{\beta(a_1, a_2)}(x) = 1 - F_{\beta(a_2, a_1)}(1 - x)$$

(поэтому, в частности, при составлении таблиц B -распределения обычно ограничиваются случаем $0 < a_1 \leq a_2$).

Основные числовые характеристики B -распределенной (с параметрами a_1 и a_2) случайной величины $\beta(a_1, a_2)$:

среднее:
$$E\beta(a_1, a_2) = \frac{a_1}{a_1 + a_2};$$

мода:
$$x_{\text{mod}} = \frac{a_1 - 1}{a_1 + a_2 - 2} \quad (\text{при } a_1 > 1 \text{ и } a_2 > 1);$$

дисперсия:
$$D\beta(a_1, a_2) = \frac{a_1 a_2}{(a_1 + a_2)^2 (a_1 + a_2 + 1)};$$

асимметрия:
$$\beta_1 = \frac{2(a_2 - a_1)\sqrt{a_1 + a_2 + 1}}{(a_1 + a_2 + 2)\sqrt{a_1 a_2}};$$

эксцесс:
$$\beta_2 = \frac{3(a_1 + a_2 + 1)[2(a_1 + a_2)^2 + a_1 a_2(a_1 + a_2 - 6)]}{a_1 a_2(a_1 + a_2 + 2)(a_1 + a_2 + 3)} - 3.$$

ВЫВОДЫ

1. *Модельный* (т. е. *аналитический*) способ задания закона распределения вероятностей (з.р.в.) является наиболее удобной формой его описания. При этом способ вычисления вероятности того, что исследуемая дискретная случайная величина ξ примет заданное значение x (т. е. способ вычисления значения $P\{\xi = x\}$), так же как и способ вычисления значения функции плотности распределения вероятностей $x(x)$ в заданной точке x для непрерывной случайной величины, *определяется явным аналитическим заданием* подходящей функции (см. формулы (3.3)–(3.10), (3.12), (3.13), (3.17')). Иногда этот способ называют *параметрической формой задания з.р.в.*

2. Следует выделить два направления использования моделей з.р.в. в теории и практике вероятностно-статистических и эконометрических исследований. *Первое направление связано с так называемым реалистическим моделированием поведения конкретных анализируемых признаков* — среднедушевого семейного дохода, тех или иных поведенческих или производственных характеристик хозяйственных субъектов, параметров финансового рынка или демографической ситуации и т. п. В этом случае модель з.р.в. призвана описать и объяснить механизм генерирования (*генезис*) анализируемых исходных данных. Другая роль широко распространенных моделей з.р.в. — *их использование в качестве вспомогательного технического средства при реализации различных методов статистического анализа данных* — при построении разного рода статистических оценок и статистических критериев (см. гл. 6–8).

3. К наиболее распространенным в практике вероятностно-статистических исследований моделям з.р.в., описывающим поведение *дискретных* случайных величин, следует отнести биномиальный, гипергеометрический, пуассоновский, полиномиальный и отрицательный биномиальный законы.

4. При описании поведения *непрерывных* случайных величин в области социально-экономических исследований наиболее распространены модели нормального (гауссовского), логарифмически-нормального, паретовского, вейбулловского, экспоненциального, лапласовского, равномерного, Коши, гамма- и бета-распределений.

5. При подборе модели «под анализируемые реальные данные» следует представлять себе общую схему (или механизм) формирования значений той или иной «модельной» случайной величины. Так, например, условия генерирования и специфика пуассоновской случайной величины определяют ее как *число наступления интересующего нас события за единицу времени, когда факт наступления такого собы-*

тия в случайном эксперименте (испытании, наблюдении) не зависит от того, сколько раз и в какие моменты времени оно наступило в прошлом, и не влияет на будущее, а испытания проводятся в стационарном режиме. Аналогично могут быть описаны и условия генерирования других модельных случайных величин.

6. Грамотная реализация основных процедур математико-статистического анализа данных, в частности, построение точечных и интервальных оценок для неизвестных значений параметров анализируемой модели (см. гл. 7), статистическая проверка гипотез, связанных со структурой и параметрами рассматриваемой модели (см. гл. 8), невозможна без знания моделей «хи-квадрат», студентовского и фишеровского F -распределений и умения пользоваться их таблицами.

7. Прикладные возможности моделей *многомерных* з.р.в. значительно скромнее. Наибольшее распространение здесь получили многомерный нормальный закон (для непрерывной многомерной случайной величины, см. (3.10)) и полиномиальный з.р.в. (для дискретной многомерной случайной величины, см. (3.7)).

ГЛАВА 4. ОСНОВНЫЕ РЕЗУЛЬТАТЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ

В гл. 2 и 3 изложены основные понятия теории вероятностей, включая набор моделей законов распределения, наиболее распространенных в теории и практике статистической обработки данных. Настоящая глава посвящена описанию некоторых связей, существующих между этими понятиями и моделями, а также отдельных их свойств, полезных для понимания сущности излагаемых далее методов экономического моделирования и статистического анализа данных.

4.1. Неравенство Чебышева

В п. 2.6.3 мы познакомились с основной характеристикой степени случайного разброса значений случайной величины — с ее дисперсией $\sigma^2 = D\xi$. Из смысла этой характеристики следует, что вероятность зафиксировать при наблюдении случайной величины ξ значение, отклоняющееся от ее среднего $a = E\xi$ не менее чем на заданную величину Δ , должна расти с ростом σ^2 . Чем больше величина дисперсии σ^2 , тем более вероятны значительные отклонения значений исследуемой случайной величины от своего центра группирования $a = E\xi$. Конечно, зная плотность (или полигон) распределения вероятностей $f_\xi(x)$, можно точно вычислить вероятность событий вида $\{|\xi - a| \geq \Delta\}$, а именно

$$P\{|\xi - a| \geq \Delta\} = \begin{cases} \int_{x: |x-a| \geq \Delta} f_\xi(x) dx, & \text{если } \xi \text{ непрерывна;} \\ \sum_{i: |x_i^0 - a| \geq \Delta} p_i, & \text{если } \xi \text{ дискретна.} \end{cases} \quad (4.1)$$

Так, например, если ξ подчиняется (a, σ^2) -нормальному закону распределения, то вероятность событий вида $|\xi - a| \geq \Delta$ зависит только от того, сколько раз в заданной величине отклонения Δ «уложится» среднее квадратическое отклонение $\sigma = \sqrt{D\xi}$.

Однако хотелось бы уметь оценивать вероятности таких событий, *опираясь только на знание величины дисперсии* $\sigma^2 = D\xi$, *не обращаясь к точному знанию закона распределения анализируемого признака* ξ . Именно эта задача и решается с помощью неравенства, выведенного русским математиком П. Л. Чебышевым¹:

$$P\{|\xi - a| \geq \Delta\} \leq \frac{\sigma^2}{\Delta^2}, \quad (4.2)$$

где $a = E\xi$ и $\sigma^2 = D\xi = E(\xi - a)^2$.

Доказательство этого неравенства несложно:

$$\begin{aligned} P\{|\xi - a| \geq \Delta\} &= \int_{x: |x-a| \geq \Delta} f_\xi(x) dx \leq \int_{x: |x-a| \geq \Delta} \left(\frac{x-a}{\Delta}\right)^2 \\ &\times f_\xi(x) dx \leq \frac{1}{\Delta^2} \int_{-\infty}^{\infty} (x-a)^2 f_\xi(x) dx = \frac{D\xi}{\Delta^2} \end{aligned}$$

(в случае дискретного признака доказательство проводится аналогично с заменой «элементов вероятностей» $f_\xi(x) dx$ вероятностями $p_i = P\{\xi = x_i^0\}$, а интегралов — соответствующими суммами).

Из хода доказательства видно, что *если распределение исследуемого признака ξ симметрично* (относительно $a = E\xi$), то имеют место и односторонние аналоги неравенства:

$$P\{\xi - a \geq \Delta\} = P\{a - \xi \geq \Delta\} \leq \frac{\sigma^2}{2\Delta^2}. \quad (4.2')$$

Как и всякий общий результат, не использующий сведения о конкретном виде распределения случайной величины ξ , неравенство Чебышева *дает лишь грубые оценки сверху* для вероятностей событий вида $\{|\xi - a| \geq \Delta\}$. Так, например, если оценивать вероятность события $\{|\xi - a| > 3\sigma\}$ для нормального признака ξ , не зная, что он подчиняется гауссовскому закону (т. е. используя неравенство Чебышева), то получим

$$P\{|\xi - a| \geq 3\sigma\} \leq \frac{\sigma^2}{(3\sigma)^2} = \frac{1}{9}.$$

¹ См.: Чебышев П. Л. О средних величинах. — Математический сборник, 1867, II. В том же 1867 г. во французском журнале «Journ. math. pures et appl», XII, была опубликована работа Бьенэмэ, также содержащая как идею доказательства этого неравенства, так и само неравенство. Поэтому неравенство (4.2) часто называют также неравенством Бьенэме-Чебышева.

Интересно сравнить эту величину с *точным значением* этой же вероятности, которое получается с помощью таблиц нормального распределения и равно 0,0027: мы видим, что точное значение вероятности в 40 (!) раз меньше ее грубой оценки, полученной на основании неравенства Чебышева.

4.2. Закон больших чисел и его следствия

Давно было замечено, что результаты отдельных наблюдений (будь то экономические, демографические, физические, метеорологические или иные наблюдения), хотя и произведенных в относительно однородных условиях, колеблются сильно, в то время как *средние из большого числа наблюдений обнаруживают замечательную устойчивость*.

Математическим основанием этого факта служат различные формы так называемого закона больших чисел. Формулировку первого частного варианта этого закона связывают с именем французского математика С. Д. Пуассона (Poisson S. D. Recherches sur la probabilité de jugements en matière criminelle et en matière civile..., Paris, Gauthier-Villars, 1837). В формулировке, приведенной ниже, этот закон был впервые доказан А. Я. Хинчиным (см. Hintchin A. Sur la loi des grands nombres. Comptes rendus de l' Académie des Sciences, 189 (1929), 477–479).

4.2.1. Закон больших чисел

Пусть $\xi_1, \xi_2, \dots, \xi_n$ — последовательность независимых и одинаково распределенных случайных величин. Если среднее значение $a = E\xi_i$ существует, то среднее арифметическое случайных величин $\xi_1, \dots, \xi_n$ по мере неограниченного роста числа слагаемых (т. е. при $n \rightarrow \infty$) *сходится по вероятности* к этому теоретическому среднему значению a , т. е. для любых сколь угодно малых положительных величин ε и δ наступает такой «момент» n_0 , начиная с которого (т. е. при всех $n \geq n_0$) будет справедливо неравенство

$$P\left\{\left|\frac{\xi_1 + \dots + \xi_n}{n} - a\right| < \varepsilon\right\} > 1 - \delta. \quad (4.3)$$

Доказательство этого утверждения не вызывает затруднений, если дополнительно потребовать существования конечной дисперсии случайных слагаемых ξ_i , т. е. существования $D\xi_i = \sigma^2 < \infty$. Действительно, в этом случае для доказательства (4.3) достаточно воспользоваться неравенством Чебышева (4.2) применительно к случайной величине

$\bar{\xi}(n) = (\xi_1 + \dots + \xi_n)/n$. Легко подсчитываются $E\bar{\xi}(n) = a$ и $D\bar{\xi}(n) = \sigma^2/n$, и, следовательно, в соответствии с (4.2)

$$P\{|\bar{\xi}(n) - a| \geq \varepsilon\} \leq \frac{D\bar{\xi}(n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Выбрав $n_0 \geq \frac{\sigma^2}{\delta\varepsilon^2}$, мы обеспечим выполнение (4.3) при любых сколь угодно малых заданных значениях ε и δ .

В качестве следствия закона больших чисел рассмотрим следующий важный результат, объясняющий эффект устойчивости *относительных частот*¹.

4.2.2. Теорема Бернулли

Пусть производится n независимых случайных экспериментов (или наблюдений случайной величины ξ), в результате каждого из которых может осуществиться или не осуществиться некоторое интересующее нас событие A (например, событие, заключающееся в том, что $\xi \in \Delta X$, где ΔX — заданная измеримая область возможных значений случайной величины ξ). Тогда при неограниченном увеличении числа экспериментов n относительная частота $\hat{p}^{(n)}(A)$ появления события A сходится по вероятности к вероятности этого события $p(A)$, т. е. для любых наперед заданных и сколь угодно малых положительных величин ε и δ наступит такой «момент» (в проведении эксперимента) n_0 , что для всех $n \geq n_0$ будет справедливо неравенство

$$P\{|\hat{p}^{(n)}(A) - p(A)| < \varepsilon\} > 1 - \delta. \quad (4.3')$$

Доказательство этого утверждения получается из (4.3), если в качестве участвующих там случайных величин ξ_i рассмотреть признаки

$$\xi_i = \begin{cases} 1, & \text{если в результате } i\text{-го эксперимента} \\ & \text{осуществилось событие } A; \\ 0 - & \text{в противном случае.} \end{cases} \quad (4.4)$$

Из определения следует, что все эти случайные величины $\xi_1, \dots, \xi_n$

¹ Теорема Бернулли исторически появилась как первая самостоятельная версия закона больших чисел намного раньше более сильной теоремы (4.3) — в 1713 г. Однако в современном изложении ее, естественно, удобнее подавать как простое следствие результата (4.3).

независимы и имеют один и тот же закон распределения, в частности:

$$\begin{aligned} E\xi_i &= 1 \cdot p(A) + 0 \cdot (1 - p(A)) = p(A); \\ D\xi_i &= (1 - p(A))^2 \cdot p(A) + (0 - p(A))^2 (1 - p(A)) \\ &= p(A)(1 - p(A)). \end{aligned} \quad (4.5)$$

Очевидно, в этом случае $\bar{\xi}(n) = (\xi_1 + \dots + \xi_n)/n$ есть не что иное, как относительная частота $\hat{p}^{(n)}(A)$ появления события A в n произведенных случайных экспериментах, причем

$$\begin{aligned} E\bar{\xi}(n) &= E\hat{p}^{(n)}(A) = p(A); \\ D\bar{\xi}(n) &= D\hat{p}^{(n)}(A) = \frac{p(A)(1 - p(A))}{n}. \end{aligned} \quad (4.6)$$

Применяя к случайным величинам (4.4) закон больших чисел (4.3), мы и получаем, с учетом (4.5) и (4.6), доказательство теоремы Бернулли (4.3').

4.3. Особая роль нормального распределения: центральная предельная теорема (ЦПТ)

Смысл результатов п. 4.2 заключается, грубо говоря, в том, что при осреднении большого числа (n) случайных слагаемых все менее ощущается характерный для случайных величин неконтролируемый разброс в их значениях, так что *в пределе* по $n \rightarrow \infty$ этот разброс исчезает вовсе или, как принято говорить, *случайная величина вырождается в неслучайную*. Однако при любом конечном числе слагаемых n случайный разброс у среднего арифметического этих слагаемых остается. Поэтому возникает вопрос исследования (опять-таки асимптотического по $n \rightarrow \infty$) характера этого разброса. Фундаментальный результат в этом направлении (известный как «центральная предельная теорема») был впервые сформулирован в упомянутом выше труде Лапласа (1812 г.) и заключается он в том, что для широкого класса независимых случайных величин $\xi_1, \dots, \xi_n$ предельный (по $n \rightarrow \infty$) закон распределения их нормированной суммы вне зависимости от типа распределения слагаемых стремится к нормальному закону распределения. Однако эта формулировка нуждается в уточнениях: что значит «нормированная» сумма случайных величин и *в каком именно смысле* закон распределения одной случайной величины стремится к закону распределения другой? Существует несколько вариантов точных

формулировок центральной предельной теоремы, отличающихся друг от друга степенью общности и видом постулируемых ограничительных условий. Мы приведем здесь формулировку Линдеберга и Леви¹.

4.3.1. Центральная предельная теорема

Если $\xi_1, \xi_2, \dots, \xi_n$ — независимые случайные величины, имеющие один и тот же закон распределения со средним значением $E\xi_i = a$ и с дисперсией $D\xi_i = \sigma^2$, то по мере неограниченного увеличения n функция распределения случайной величины

$$\bar{\xi}^*(n) = \frac{\frac{1}{n}(\xi_1 + \dots + \xi_n) - a}{\sigma/\sqrt{n}} = \frac{(\bar{\xi}(n) - a)\sqrt{n}}{\sigma}$$

стремится к функции распределения стандартного нормального закона при любом заданном значении их аргументов, т. е.

$$F_{\bar{\xi}^*(n)}(x) \rightarrow \Phi(x) \quad \text{при } n \rightarrow \infty$$

для любого значения x , где $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$.

Таким образом, центральная предельная теорема дает математически строгое описание условий, индуцирующих механизм нормального закона распределения (см. неформальное обсуждение этих условий в п. 3.1.5). Она оправдывает, в частности, закономерность той центральной роли, которую играет нормальное распределение в теории и практике статистических исследований. Содержание центральной предельной теоремы можно считать дальнейшим (после закона больших чисел) уточнением стохастического поведения среднего арифметического из ряда случайных величин.

Центральная предельная теорема может быть распространена в различных направлениях: когда случайные слагаемые не являются одинаково распределенными (формулировка А. М. Ляпунова); когда компоненты ξ_i не являются независимыми; наконец, когда случайные величины ξ_i являются многомерными.

¹ L é v y P. Calcul des probabilités. Paris, 1925; Lindeberg J. W. Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. — Math. Zeitschr., 15 (1922), 211.

4.3.2. Многомерная центральная предельная теорема

Пусть $\xi_1, \xi_2, \dots, \xi_n$ — независимые и одинаково распределенные p -мерные случайные величины с вектором средних значений $M = E\xi_i = (E\xi_i^{(1)}, E\xi_i^{(2)}, \dots, E\xi_i^{(p)})^T$ и ковариационной матрицей $\Sigma = E\{(\xi_i - M) \times (\xi_i - M)^T\}$. Тогда при $n \rightarrow \infty$ совместная функция распределения случайного вектора $\xi^*(n) = [\sum_{i=1}^n (\xi_i - M)]/\sqrt{n}$ сходится (для любого значения векторного аргумента X) к совместной функции распределения p -мерной нормальной случайной величины, имеющей вектор средних $0 = (0, 0, \dots, 0)'$ и ковариационную матрицу Σ .

4.3.3. Комментарии к центральной предельной теореме

З а м е ч а н и е 1. Необходима известная осторожность при использовании центральной предельной теоремы в практике статистических исследований.

Во-первых, если *предельный* вид распределения суммы случайных слагаемых при определенных условиях всегда нормален и не зависит от вида распределения самих слагаемых, то *скорость сходимости* распределения суммы к нормальному закону существенно зависит от типа распределения исходных компонент. Так, например, при суммировании равномерно распределенных случайных величин уже при 6–10 слагаемых можно добиться достаточной близости к нормальному закону, в то время как для достижения той же близости при суммировании χ^2 -распределенных слагаемых понадобится более 100 слагаемых.

Во-вторых, центральной предельной теоремой вообще не рекомендуется пользоваться для аппроксимации вероятностей на «хвостах» распределения, т. е. при оценке вероятностей событий вида $\{\bar{\xi}^{(*)}(n) < \tilde{x}_{\min}\}$ и $\{\bar{\xi}^{(*)}(n) \geq \tilde{x}_{\max}\}$, где $\tilde{x}_{\min}$ и $\tilde{x}_{\max}$ — возможные значения, близкие соответственно к левой и правой границам диапазона изменения исследуемого признака $\bar{\xi}^{(*)}(n)$. Поскольку в этом случае числовые значения вероятностей $P\{\bar{\xi}^{(*)}(n) < \tilde{x}_{\min}\} = F_{\bar{\xi}^{(*)}(n)}(\tilde{x}_{\min})$ и $P\{\bar{\xi}^{(*)}(n) > \tilde{x}_{\max}\} = 1 - F_{\bar{\xi}^{(*)}(n)}(\tilde{x}_{\max})$ малы, то из малости разностей $F_{\bar{\xi}^{(*)}(n)}(\tilde{x}_{\min}) - \Phi(\tilde{x}_{\min})$ и $F_{\bar{\xi}^{(*)}(n)}(\tilde{x}_{\max}) - \Phi(\tilde{x}_{\max})$ (которая вытекает из центральной предельной теоремы) вовсе не следует малость *относительных ошибок аппроксимации*

$$\frac{F_{\bar{\xi}^{(*)}(n)}(\tilde{x}_{\min})}{\Phi(\tilde{x}_{\min})} \quad \text{и} \quad \frac{1 - F_{\bar{\xi}^{(*)}(n)}(\tilde{x}_{\max})}{1 - \Phi(\tilde{x}_{\max})},$$

которые, как правило, оказываются чрезмерно большими. Так, например, пусть $\bar{\xi}^{(*)}(n)$ — нормированный среднедушевой доход в семье (соответ-

ственно $\xi_1, \xi_2, \dots$ — заработная плата работающих членов семьи и другие составляющие семейного дохода) и пусть нас интересует доля q семей с очень высоким доходом, а именно с доходом, не меньшим некоторого достаточно высокого уровня $\tilde{x}_{\max}$ (руб.). Исследования показали, что точное значение этой доли $q = 1 - F_{\xi^*(n)}(\tilde{x}_{\max}) = 0,03$, в то время как соответствующая нормальная аппроксимация дала результат $\hat{q} = 1 - \Phi(\tilde{x}_{\max}) = 0,003$. Разность $q - \hat{q}$ сама по себе мала (как и следует из центральной предельной теоремы), однако относительная погрешность нормальной аппроксимации в данном случае составляет десятикратную величину, т. е. 1000%! Особенно важным это предостережение оказывается при попытках использования нормальных аппроксимаций в задачах расчета зависимостей типа «предельная прочность (или пропускная способность) системы — вероятность разрушения (отказа в обслуживании)».

З а м е ч а н и е 2. Центральная предельная теорема позволяет проследить асимптотические связи, существующие между различными модельными законами распределения (см. гл. 3), с одной стороны, и нормальным законом — с другой. Опираясь на центральную предельную теорему, можно объяснить, в частности, следующие полезные для статистической практики факты:

1. Распределение (n, p) -биномиальной случайной величины $\xi(p, n)$ асимптотически (по $n \rightarrow \infty$) нормально¹ с параметрами $E\xi(p, n) = np$ и $D\xi(p, n) = np(1 - p)$. Данный результат известен как *теорема Муавра-Лапласа* (доказана впервые Муавром в 1733 г., когда еще не была известна центральная предельная теорема) и является прямым следствием центральной предельной теоремы, примененной к случайным величинам (4.4) с учетом (4.5).

2. Распределение λ пуассоновской случайной величины $\xi(\lambda)$ асимптотически (по $\lambda \rightarrow \infty$) нормально с параметрами $E\xi(\lambda) = \lambda$ и $D\xi(\lambda) = \lambda$.

3. Распределение (n, N, M) -гипергеометрической случайной величины $\xi(N, M, n)$ асимптотически (по $N \rightarrow \infty, M \rightarrow \infty, \frac{M}{N} \rightarrow p > 0$ и $n \rightarrow \infty$) нормально с параметрами $E\xi(N, M, n) = np$ и $D\xi(N, M, n) = np(1 - p)$.

4. Функция распределения нормированной k -мерной $(p_1, p_2, \dots, p_k; n)$

¹ Случайная величина $\xi(n)$, зависящая от параметра n , называется *асимптотически (по n) нормальной*, если существуют такие нормирующие не случайные переменные $A(n)$ и $B(n)$, что функция распределения случайной величины $\eta(n) = A(n) \cdot \xi(n) + B(n)$ стремится при $n \rightarrow \infty$ к функции распределения стандартного нормального закона при любом заданном значении их аргумента x .

полиномиальной случайной величины

$$\xi = \left(\frac{\xi^{(1)} - np_1}{\sqrt{np_1}}, \dots, \frac{\xi^{(k)} - np_k}{\sqrt{np_k}} \right)$$

при $n \rightarrow \infty$ стремится к функции распределения несобственного (вырожденного) k -мерного нормального закона с вектором нулевых средних значений и с ковариационной матрицей

$$\Sigma = I_k - \begin{pmatrix} \sqrt{p_1} \\ \vdots \\ \sqrt{p_k} \end{pmatrix} (\sqrt{p_1}, \sqrt{p_2}, \dots, \sqrt{p_k}),$$

имеющей ранг, равный $k - 1$ ¹.

5. Распределение случайной величины $\chi^2(m)$ асимптотически (по $m \rightarrow \infty$) нормально с параметрами $E\chi^2(m) = m$ и $D\chi^2(m) = 2m$.

6. Распределение случайной величины $t(m)$ асимптотически (по $m \rightarrow \infty$) нормально с параметрами $Et(m) = 0$ и $Dt(m) = 1$.

4.4. Законы распределения вероятностей случайных признаков, являющихся функциями от известных случайных величин

В теории и практике статистических исследований очень важно уметь вычислять закон распределения вероятностей для функций от случайных величин, распределение которых известно. Именно на этом, главным образом, основана теория статистического оценивания и проверки статистических гипотез (см. гл. 7 и 8), так как и *статистическая оценка*, и *критическая статистика*, используемые соответственно при оценивании неизвестных значений параметров и при построении критериев статистической проверки гипотез, суть не что иное, как *функции от результатов наблюдения* исследуемой случайной величины ξ . Для того чтобы ими осмысленно пользоваться и знать их статистические свойства, мы должны уметь восстанавливать их закон распределения по распределению изучаемой случайной величины ξ (а значит, и ее наблюдений). Ниже

¹ Здесь и далее символ I_k означает единичную матрицу размерности k .

описываются основные правила, руководствуясь которыми можно решать эту задачу.

1. Пусть случайная величина η является *монотонно возрастающей непрерывной функцией* от заданной непрерывной случайной величины ξ , имеющей всюду дифференцируемую функцию распределения $F_\xi(x)$, т. е. $\eta = g(\xi)$. Каждому возможному значению x случайной величины ξ будет соответствовать возможное значение $y = g(x)$ случайной величины η .

В силу монотонности и непрерывности преобразования g по заданному значению η можно однозначно восстановить соответствующее ξ с помощью преобразования, обратного к g (обозначим его g^{-1}), т. е. $\xi = g^{-1}(\eta)$. Аналогичное соотношение связывает и возможные значения этих случайных величин, т. е. $x = g^{-1}(y)$. Заметим, что функция $g^{-1}(y)$, так же как и функция $g(x)$, является монотонно возрастающей непрерывной функцией.

Попробуем выразить функцию распределения $F_\eta(y)$ в терминах заданных нам функций $F_\xi(x)$, g и g^{-1} .

$$\begin{aligned} F_\eta(y) &= P\{\eta < y\} = P\{g^{-1}(\eta) < g^{-1}(y)\} \\ &= P\{\xi < g^{-1}(y)\} = F_\xi(g^{-1}(y)). \end{aligned} \quad (4.7)$$

Дифференцирование обеих частей (4.7) по y дает

$$f_\eta(y) = f_\xi(g^{-1}(y)) \cdot \frac{dg^{-1}(y)}{dy}. \quad (4.8)$$

Точно такие же рассуждения в случае *монотонно убывающей функции* $g(x)$ приведут нас к некоторому видоизменению формулы (4.8):

$$f_\eta(y) = f_\xi(g^{-1}(y)) \cdot \left(- \frac{dg^{-1}(y)}{dy} \right). \quad (4.8')$$

Можно объединить формулы (4.8) и (4.8') в одной, справедливой для любого взаимно-однозначного преобразования g :

$$f_\eta(y) = f_\xi(g^{-1}(y)) \cdot \left| \frac{dg^{-1}(y)}{dy} \right|. \quad (4.8'')$$

Пример 4.1. Вычислить функцию плотности случайной величины $\eta = e^\xi$, если известно, что ξ подчиняется (a, σ^2) — нормальному закону. В данном примере $g(x) = e^x$, $g^{-1}(y) = \ln y$, следовательно,

$$\frac{dg^{-1}(y)}{dy} = \frac{d(\ln y)}{dy} = \frac{1}{y}.$$

Подставляя это в (4.8''), получаем

$$f_{\eta}(y) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(\ln y - a)^2}{2\sigma^2}} \cdot \frac{1}{y},$$

т. е. плотность *логарифмически-нормального закона* (см. п. 3.1.6).

Пример 4.2. Вычислить функцию плотности случайной величины $\eta = a + b\xi$, если известна плотность $f_{\xi}(x)$ случайной величины ξ . В данном случае $g(x) = a + bx$, $g^{-1}(y) = (y - a)/b$ и $(dg^{-1}(y)/dy) = 1/b$. В соответствии с (4.7) и (4.8'') имеем:

$$\begin{aligned} F_{\eta}(y) &= F_{\xi}\left(\frac{y-a}{b}\right); \\ f_{\eta}(y) &= \left(\frac{y-a}{b}\right) \cdot \left|\frac{1}{b}\right|. \end{aligned} \quad (4.9)$$

Это правило пересчета функций распределения и плотности позволяет, в частности, использовать таблицы *стандартного*, т. е. (0; 1)-нормального закона для определения значений функции распределения и функции плотности нормальной случайной величины $\xi(a, \sigma^2)$ с произвольными параметрами $a = E\xi(a, \sigma^2)$ и $\sigma^2 = D\xi(a, \sigma^2)$. При этом, как легко видеть, роль ξ играет стандартная нормальная величина $\xi(0, 1)$, а роль η — произвольная нормальная величина $\xi(a, \sigma^2)$, т. е.

$$\xi(a, \sigma^2) = a + \sigma \cdot \xi(0, 1). \quad (4.10)$$

Соответственно имеем:

$$\begin{aligned} \Phi(x; a, \sigma^2) &= \Phi\left(\frac{x-a}{\sigma}; 0, 1\right) = \Phi\left(\frac{x-a}{\sigma}\right); \\ \varphi_{\xi(a, \sigma^2)}(x) &= \frac{1}{\sigma} \varphi_{\xi(0, 1)}\left(\frac{x-a}{\sigma}\right). \end{aligned} \quad (4.9')$$

2. Если интересующее нас преобразование $\eta = g(\xi)$ не является взаимно-однозначными, то сколько-нибудь общие формулы получить не удастся. Вместо этого приходится каждый раз решать определенный тип задач, прилаживаясь к их специфике. Рассмотрим, например, случай $\eta = \xi^2$:

$$\begin{aligned} F_{\eta}(y) &= P\{\eta < y\} = P\{|\xi| < \sqrt{y}\} = P\{-\sqrt{y} < \xi < \sqrt{y}\} \\ &= F_{\xi}(\sqrt{y}) - F_{\xi}(-\sqrt{y}). \end{aligned}$$

Следовательно, после дифференцирования левой и правой частей по y имеем:

$$\begin{aligned} f_{\eta}(y) &= f_{\xi}(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} + f_{\xi}(-\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} \\ &= \frac{1}{2\sqrt{y}} (f_{\xi}(\sqrt{y}) + f_{\xi}(-\sqrt{y})). \end{aligned}$$

Применение данной формулы к случаю, когда ξ является стандартной нормальной величиной, дает

$$f_{\eta}(y) = \frac{1}{2\sqrt{y}} \cdot 2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}} = \frac{(\frac{1}{2})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} y^{\frac{1}{2}-1} e^{-\frac{1}{2}y},$$

что является функцией плотности χ^2 -распределения с одной степенью свободы и одновременно — частным случаем гамма-распределения с параметрами $a = b = 1/2$ (см. п. 3.2.5).

3. Обобщим формулу (4.8'') на *многомерный случай*. Пусть $\xi = (\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(p)})$ — p -мерная случайная величина с известной функцией распределения $F_{\xi}(X)$ и плотностью вероятности $f_{\xi}(X)$ и пусть другая p -мерная случайная величина $\eta = (\eta^{(1)}, \dots, \eta^{(p)})$ определяется как заданная непрерывная *векторная* функция $g(\xi)$ от компонент ξ , т. е.

$$\begin{aligned} \eta^{(1)} &= g_1(\xi^{(1)}, \dots, \xi^{(p)}); \\ &\dots\dots\dots \\ \eta^{(p)} &= g_p(\xi^{(1)}, \dots, \xi^{(p)}). \end{aligned}$$

Предполагается, что соответствие $\eta = g(\xi)$ является взаимно-однозначным, т. е. существует обратное преобразование g^{-1} , позволяющее по заданному «значению» η однозначно восстанавливать соответствующее «значение» ξ :

$$\begin{aligned} \xi^{(1)} &= g_1^{-1}(\eta^{(1)}, \dots, \eta^{(p)}); \\ &\dots\dots\dots \\ \xi^{(p)} &= g_p^{-1}(\eta^{(1)}, \dots, \eta^{(p)}). \end{aligned}$$

Соответственно между многомерными возможными «значениями» $X = (x^{(1)}, \dots, x^{(p)})$ и $Y = (y^{(1)}, \dots, y^{(p)})$ случайных величин ξ и η имеют место векторные соотношения

$$Y = g(X) \quad \text{и} \quad X = g^{-1}(Y).$$

Тогда совместная плотность вероятности случайных величин $\eta = (\eta^{(1)}, \dots, \eta^{(p)})$ равна

$$f_{\eta}(y^{(1)}, \dots, y^{(p)}) = f_{\xi}(g_1^{-1}(Y), g_2^{-1}(Y), \dots, g_p^{-1}(Y)) \cdot |J(y)|, \quad (4.11)$$

где определитель преобразования (*якобиан*)

$$J(Y) = \begin{vmatrix} \frac{\partial g_1^{-1}(Y)}{\partial y^{(1)}} & \frac{\partial g_1^{-1}(Y)}{\partial y^{(2)}} & \dots & \frac{\partial g_1^{-1}(Y)}{\partial y^{(p)}} \\ \frac{\partial g_2^{-1}(Y)}{\partial y^{(1)}} & \frac{\partial g_2^{-1}(Y)}{\partial y^{(2)}} & \dots & \frac{\partial g_2^{-1}(Y)}{\partial y^{(p)}} \\ \dots & \dots & \dots & \dots \\ \frac{\partial g_p^{-1}(Y)}{\partial y^{(1)}} & \frac{\partial g_p^{-1}(Y)}{\partial y^{(2)}} & \dots & \frac{\partial g_p^{-1}(Y)}{\partial y^{(p)}} \end{vmatrix}$$

так же, как и в формуле (4.8''), берется по абсолютной величине.

4. Выведем распределение суммы двух независимых случайных слагаемых (*формулу композиции*). Пусть независимые случайные величины ξ_1 и ξ_2 имеют плотности вероятности соответственно $f_{\xi_1}(x)$ и $f_{\xi_2}(y)$. Требуется произвести композицию этих плотностей, т. е. найти плотность распределения случайной величины $\eta = \xi_1 + \xi_2$. По существу, мы должны рассмотреть совместное двумерное распределение $f_{(\xi_1, \xi_2)}(x, y)$ и для определения функции распределения случайной величины η найти в плоскости xOy область возможных значений (ξ_1, ξ_2) , соответствующих событию $\{\eta < z\}$. На рис. 4.1 эта область заштрихована и обозначена A_z . Получаем

$$\begin{aligned} F_{\eta}(z) &= P\{\eta < z\} = P\{\xi_1 + \xi_2 < z\} = P\{(\xi_1, \xi_2) \in A_z\} \\ &= \int \int_{A_z} f_{(\xi_1, \xi_2)}(x, y) \, dx dy = \int \int_{A_z} f_{\xi_1}(x) \cdot f_{\xi_2}(y) \, dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_{\xi_1}(x) f_{\xi_2}(y) \, dx dy = \int_{-\infty}^{\infty} f_{\xi_1}(x) \left(\int_{-\infty}^{z-x} f_{\xi_2}(y) \, dy \right) dx. \quad (4.12) \end{aligned}$$

Здесь мы воспользовались тождеством $f_{(\xi_1, \xi_2)}(x, y) = f_{\xi_1}(x) \cdot f_{\xi_2}(y)$ (справедливым в силу независимости ξ_1 и ξ_2), а при интегрировании по области A_z пределы интегрирования по оси Ox брали от $-\infty$ до $+\infty$, а по оси Oy — от $-\infty$ до прямой $y = z - x$.

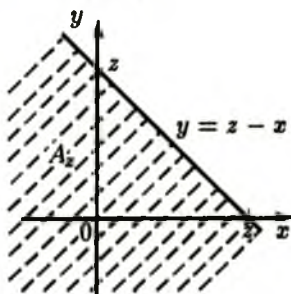


Рис. 4.1. Попадание в область A_z на плоскости xOy соответствует событию $\{\xi_1 + \xi_2 < z\}$

Дифференцирование по z левой и правой частей соотношения (4.12) дает

$$f_\eta(z) = \int_{-\infty}^{\infty} f_{\xi_1}(x) f_{\xi_2}(z-x) dx. \quad (4.13)$$

Формулу (4.13) называют формулой композиции двух распределений или *формулой свертки*. Для обозначения композиции (свертки) законов распределения часто применяют символическую запись

$$f_\eta = f_{\xi_1} * f_{\xi_2}.$$

Воспользовавшись формулой (4.13), можно вывести упомянутое в п. 3.1.5 и 3.1.10 свойство «самовоспроизводимости» законов Гаусса и Коши (сумма нормальных случайных величин сама распределена по нормальному закону; сумма одинаково по Коши распределенных случайных величин сама подчиняется закону распределения Коши), а также получить формулы для плотности распределения сумм равномерно распределенных величин, приведенные в п. 3.1.7.

ВЫВОДЫ

1. Следует выделить три типа основных результатов теории вероятностей:

доасимптотические, позволяющие анализировать основные закономерности в поведении случайной величины по ее главным числовым характеристикам — среднему значению, дисперсии и т. п. — без апелляции к знанию общего вида закона ее распределения;

асимптотические, позволяющие анализировать основные закономерности в поведении сумм большого числа случайных слагаемых (а именно устанавливать их асимптотическую устойчивость, т.е. их сходимости к некоторым постоянным значениям по мере роста числа слагаемых, или описывать асимптотический вид закона распределения этих сумм) без точного знания законов распределения отдельных слагаемых;

относящиеся к *теории преобразований случайных величин*, позволяющие находить закон распределения интересующих нас функций от набора случайных величин, совместное распределение которых нам известно.

2. Первый тип результатов представлен в данном учебнике *неравенством Чебышева*, позволяющим оценивать вероятность зафиксировать при наблюдении случайной величины ξ значение, отклоняющееся от ее среднего значения $a = E\xi$ не менее чем на заданную величину Δ , без знания закона распределения исследуемой случайной величины ξ (см. (4.2)).

3. *Закон больших чисел* и его следствия относятся к первому (менее глубокому) уровню асимптотических результатов и позволяют устанавливать сходимость (по вероятности) нормированных сумм большого числа случайных слагаемых к некоторым постоянным значениям — по мере роста числа слагаемых — практически независимо от вида распределения самих слагаемых.

4. *Центральная предельная теорема* относится к следующему (более глубокому) уровню асимптотических результатов. Она утверждает, в частности, что закон распределения нормированной суммы большого числа случайных слагаемых практически вне зависимости от типа распределения самих слагаемых стремится (по мере роста числа слагаемых) к нормальному (гауссовскому) закону распределения (см. п. 4.3.1).

Однако необходима известная осторожность при практическом использовании центральной предельной теоремы: во-первых, часто можно подобрать весьма простые (и более точные, чем нормальные!) аппроксимации для распределения суммы *небольшого конечного* числа случайных слагаемых; во-вторых, центральная предельная теорема *плохо работает на «хвостат» распределений*, т.е. при оценке вероятностей больших отклонений анализируемой суммы случайных величин от своего среднего значения.

5. Основным результатом *теории преобразования случайных величин* является правило (4.8") (или (4.11) в многомерном случае), позволяющее вычислять закон распределения (функцию плотности или полигон частот) случайной величины, являющейся заданной функцией от набора исходных случайных величин, совместный закон распределения которых нам известен.

ГЛАВА 5. ЦЕПИ МАРКОВА

5.1. Последовательности случайных экспериментов и случайных величин в дискретном вероятностном пространстве

Рассмотрим дискретное вероятностное пространство с *конечным* числом элементарных исходов $\omega_1, \omega_2, \dots, \omega_N$ и вероятностями их появления (в результате реализации случайного эксперимента), соответственно $p_1 = P\{\omega_1\}$, $p_2 = P\{\omega_2\}, \dots, p_N = P\{\omega_N\}$. Пусть нас интересует вероятность появления заданной последовательности элементарных исходов $\omega_{i_1}, \omega_{i_2}, \dots, \omega_{i_n}$ в результате реализации n последовательных случайных экспериментов. Если случайные эксперименты производятся в рамках одного и того же вероятностного пространства и *независимы друг от друга*, то мы имеем дело с **последовательностью независимых случайных экспериментов**, и интересующая нас вероятность определится по формуле вероятности независимых событий (1.13), что в нашем случае дает

$$P\{\omega_{i_1} \omega_{i_2} \dots \omega_{i_n}\} = P\{\omega_{i_1}\} \cdot P\{\omega_{i_2}\} \cdot \dots \cdot P\{\omega_{i_n}\} = p_{i_1} p_{i_2} \cdot \dots \cdot p_{i_n}. \quad (5.1)$$

Из независимости рассматриваемых событий следует также, что если бы нас интересовал только результат, полученный на одном каком-то «шаге» этой последовательности (например, на шаге t), то вероятность $P\{\omega_{i_t}\}$ определяется единственно величиной p_{i_t} и *никак не зависит от предыстории*, т. е. от того, какие именно результаты мы имели на предыдущих шагах этой последовательности. Очевидно, в рамки подобной схемы укладываются ранее рассмотренные примеры 1.1 (последовательность подбрасываний одной и той же монеты), 1.3 (четыrehкратное бросание игральной кости), 1.5 (проверка по альтернативному признаку n изделий, случайно отобранных из продукции стационарно функционирующего производства).

Однако нетрудно представить себе реальные ситуации, в которых результат случайного эксперимента на определенном шаге последовательности в определенном смысле зависит от того, какие именно результаты наблюдались на предыдущих шагах этой последовательности. Очевидно, в подобных ситуациях речь следует вести о **последовательности зависимых случайных экспериментов**.

В дальнейшем несколько видоизменим терминологию с целью ее более естественного приспособления к реальным задачам. А именно будем интерпретировать каждый из элементарных исходов ω ; как «*возможное состояние*» анализируемого объекта (индивидуума, работника, семьи, домашнего хозяйства и т. п.). Введем в рассмотрение *дискретное время* t и случайную величину $\xi_t(\omega)$, значения которой $(1, 2, \dots, N)$ определяют, в каком именно состоянии находился анализируемый объект в момент времени t ($t = 1, 2, \dots, n$), т.е. $\xi_t(\omega_i) = i$. В каждый следующий момент времени $t + 1$ объект из состояния ω_i , в котором он находился, может перейти в любое другое состояние ω_j или остаться в том же самом состоянии ω_i . Таким образом, от рассмотрения последовательностей случайных экспериментов мы переходим к рассмотрению последовательности случайных величин $\xi_1(\omega), \xi_2(\omega), \dots, \xi_n(\omega)$, *вообще говоря, взаимно-зависимых*. А чтобы «задать поведение» многомерной случайной величины $\xi = (\xi_1(\omega), \xi_2(\omega), \dots, \xi_n(\omega))$, необходимо, как мы знаем (см. выше, п. 2.5.3), задать соответствующий совместный n -мерный закон распределения вероятностей, который в данном случае (с учетом того, что каждая из компонент ξ может принимать одно из N возможных значений) описывается с помощью N^n конкретных чисел (вероятностей).

Правда, в случае последовательности *независимых* случайных экспериментов мы имеем соответствующую **последовательность независимых одинаково распределенных случайных величин**, так что в соответствии с (2.14') и (5.1)

$$P\{\xi_1 = i_1, \xi_2 = i_2, \dots, \xi_n = i_n\} = p_{i_1} \cdot p_{i_2} \cdot \dots \cdot p_{i_n} \quad (5.2)$$

и *распределение n -мерной случайной величины $\xi = (\xi_1, \xi_2, \dots, \xi_n)$ задается всего лишь N числами: $p_1, p_2, \dots, p_N$.*

Можно ожидать, что и при некоторых *специальных типах зависимостей*, существующих между элементами последовательности $\xi_1, \xi_2, \dots, \xi_n$, способ задания ее поведения окажется более простым и экономичным, чем в самом общем случае. Именно к рассмотрению одного из таких специальных типов зависимостей мы и переходим.

5.2. Последовательности, образующие цепь Маркова (определения, примеры, прикладные задачи)

Продолжая рассмотрение системы *«объект — его состояние в заданный момент времени t »*, предположим, что эта система характеризуется *«ограниченной памятью»*, а именно вероятность того, что объект в момент времени t окажется в заданном состоянии *« j »*, зависит не от всей предыстории, а *только от того, в каком состоянии он находился в предыдущий момент времени $t - 1$* , т. е.

$$P\{\xi_t = j \mid \xi_1 = i_1, \xi_2 = i_2, \dots, \xi_{t-1} = i\} = P\{\xi_t = j \mid \xi_{t-1} = i\}. \quad (5.3)$$

О таких зависимостях последовательности случайных величин $\xi_1, \xi_2, \dots, \xi_t$ говорят как о *зависимостях марковского типа*, а соответствующие последовательности называют цепями Маркова. Если дополнительно предположить, что вероятности $p_{ij}^{(1)} = P\{\xi_t = j \mid \xi_{t-1} = i\}$ *остаются одними и теми же при всех t* (т. е. не зависят от t), то соответствующая марковская цепь называется *однородной*. Именно однородным цепям Маркова и посвящена данная глава.

Чтобы дать строгое определение однородной цепи Маркова, необходимо задать способ вычисления совместного распределения вероятностей для многомерных случайных величин $(\xi_1, \xi_2, \dots, \xi_t)$ при любых $t = 1, 2, \dots$.

О п р е д е л е н и е. *Последовательность случайных величин $\xi_1, \xi_2, \dots$, каждая из которых может принимать конечное число значений $x_1^0, x_2^0, \dots, x_N^0$, называется однородной цепью Маркова, если вероятности событий вида*

$$\{\xi_1 = x_{i_1}^0, \xi_2 = x_{i_2}^0, \dots, \xi_t = x_{i_t}^0\} \quad (5.*)$$

определяются формулой

$$P\{\xi_1 = x_{i_1}^0, \xi_2 = x_{i_2}^0, \dots, \xi_t = x_{i_t}^0\} = p_{i_1} \cdot p_{i_1 i_2} \cdot \dots \cdot p_{i_{t-1} i_t} \quad (5.4)$$

через начальное распределение вероятностей

$$p_1 = P\{\xi_1 = x_1^0\}, p_2 = P\{\xi_1 = x_2^0\}, \dots, p_N = P\{\xi_1 = x_N^0\} \quad (5.5)$$

¹ Требование конечности числа возможных состояний не обязательно при определении цепи Маркова. На самом деле допускается и счетное число состояний. Однако в рамках наших целей можно обойтись и конечным числом состояний.

(очевидно, при этом $p_{00} = 1 - p_{01}$ и $p_{10} = 1 - p_{11}$). А. А. Марков показал, как, располагая, помимо значений вероятностей p_{01} и p_{11} , еще величиной вероятности появления события $\{\xi = 1\}$ в первом испытании, можно вычислить вероятность любого события вида (5.*). Он назвал эту схему **простейшей цепью**. В дальнейшем он обобщил эту схему на случай произвольного конечного или даже счетного числа возможных значений случайной величины. Это обобщение и получило название цепи Маркова.

Пример 5.1. Динамика типологии семейного потребления. В данном примере исследуемыми *объектами* являются семьи, точнее, домашние хозяйства. *Возможные «состояния»* объекта — типы потребительского поведения. Единица измерения (такт) времени — один год. Соответственно, элемент p_i *начального распределения вероятностей* интерпретируется как доля домашних хозяйств, относящихся к i -му типу потребительского поведения в начальный (первый) «момент времени», а *переходная вероятность* p_{ij} — как доля семей i -го типа потребительского поведения, переходящих за 1 год в «состояние» j .

Пример 5.2. Игра двух лиц до разорения. Исследуемый *объект* — один из двух участников игры, которая происходит по следующим правилам. Суммарный капитал двух игроков фиксирован и равен N денежным единицам. Выигрыш одного кона составляет одну денежную единицу. Анализируемый игрок начинает игру, имея n единиц, и выигрывает кон с вероятностью p . После проигрыша последней единицы i -м игроком ($i = 1, 2$) игра заканчивается его разорением. Очевидно, «состояние» объекта в этой схеме следует понимать в буквальном смысле: оно определяется имеющейся у него в наличии суммой денежных единиц в соответствующий «момент времени», а последний интерпретируется как очередной кон игры. Итак, мы имеем $N + 1$ возможных состояний, а именно: $0; 1; 2; \dots; N$.

Начальное распределение вероятностей в данном случае определяется формулой

$$p_i = P\{\xi(1) = i\} = \begin{cases} 1 & \text{при } i = n; \\ 0 & \text{для всех } i \neq n. \end{cases}$$

В соответствии с правилами игры *переходные вероятности* p_{ij} определяются соотношениями

$$p_{ij} = \begin{cases} p & \text{при } j - i = 1 \text{ и } i \geq 1; \\ 1 - p & \text{при } i - j = 1 \text{ и } i \geq 1; \\ 1 & \text{при } i = j = 0 \text{ и при } i = j = N. \end{cases}$$

Пример 5.3. Демографическая модель передвижки возрастов. В качестве *объекта* рассматривается элемент населения страны или региона (индивидуум), а его «состояние» определяется его возрастом. Смежные «состояния-возрасты» отличаются друг от друга на один

год. Исключение составляет состояние, определяемое как «предельный возраст»: к нему будем относить все индивидуумы в возрасте, превышающим 85 лет. Кроме того, вводятся еще три особых состояния: «иммиграционное», «эмиграционное» и «поглощающее». Иммиграционное и эмиграционное состояния связаны с возможным соответственно прибытием и выбытием индивидуумов из анализируемого региона в связи с миграцией населения; а поглощающее — только с выбытием по причине смерти. Таким образом, мы имеем 89 возможных состояний: первое — возраст до 1 года, второе — до 2 лет, ..., 85-е — до 85 лет, 86-е — свыше 85 лет, 87-е — иммиграционное, 88-е — эмиграционное и 89-е — поглощающее.

Начальное распределение вероятностей $p_i = P\{\xi(1) = i\}$, $i = 1, 2, \dots, 86, 87, 88, 89$ определяется демографической структурой общества, зафиксированной на конец первого года исследуемого периода. При этом под p_{87} понимается доля иммигрантов этого года во всем населении региона, а p_{88} — доля эмигрантов. Матрица P переходных вероятностей p_{ij} может быть определена в терминах возрастных характеристик смертности (μ_i), иммиграционной (q_i) и эмиграционной (β_i) активности населения в виде таблицы 5.1.

Таблица 5.1

Но- мер сос- тоя- ния (<i>i</i>)	Номер состояния (<i>j</i>)										
	1	2	3	4	5	...	85	86	87	88	89
1	0	$1 - \mu_1 - \beta_1$	0	0	0	...	0	0	0	β_1	μ_1
2	0	0	$1 - \mu_2 - \beta_2$	0	0	...	0	0	0	β_2	μ_2
3	0	0	0	$1 - \mu_3 - \beta_3$	0	...	0	0	0	β_3	μ_3
⋮	⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮	⋮	⋮
85	0	0	0	0	0	...	0	$1 - \mu_{85} - \beta_{85}$	0	β_{85}	μ_{85}
86	0	0	0	0	0	...	0	$1 - \mu_{86} - \beta_{86}$	0	β_{86}	μ_{86}
87	q_1	q_2	q_3	q_4	q_5	...	q_{85}	q_{86}	0	0	0
88	0	0	0	0	0	...	0	0	0	1	0
89	0	0	0	0	0	...	0	0	0	0	1

Участвующие в вычислении вероятностей p_{ij} возрастные демографические характеристики интерпретируются следующим образом. Величина μ_i оценивается как доля тех людей возраста i , которые умерли в этом возрасте; значение β_i определяет долю тех людей возраста i , которые эми-

грировали из анализируемого региона (страны) в данном возрасте; наконец, q_i выражает удельный вес людей i -го возраста среди всех иммигрантов данного региона, т. е. среди всех людей, прибывающих в анализируемый регион на постоянное поселение.

Пример 5.4. Модель пребывания студента в вузе. В качестве объекта рассматривается студент вуза с пятилетним сроком обучения. Его i -е состояние соответствует его обучению на i -м курсе ($i = 1, 2, \dots, 5$); 6-е состояние — специалист, окончивший данный вуз; 7-е — лицо, обучавшееся в данном вузе, но по каким-то причинам не окончившее его. Чтобы не увеличивать числа возможных состояний примем допущение, что по правилам данного вуза зачисление в него переводом студентов из других вузов невозможно. Из смысла состояний следует, что ненулевыми элементами матрицы P переходных вероятностей могут быть только величины $p_{11}, p_{12}, p_{17}, p_{22}, p_{23}, p_{27}, p_{33}, p_{34}, p_{37}, p_{44}, p_{45}, p_{47}, p_{55}, p_{56}, p_{57}$, а также $p_{66} = p_{77} = 1$.

Пример 5.5. Игровые возможности двух шахматистов. Объектом в данном примере является соревнующийся шахматист, а состоянием — результат сыгранной им партии (1-е состояние — выигрыш, 2-е состояние — ничья, 3-е состояние — поражение). Анализируется турнирное поведение двух шахматистов A и B . Игрок A каждую партию турнира независимо от исходов предыдущих партий выигрывает с вероятностью «в», делает ничью с вероятностью «н» и проигрывает с вероятностью «п» $= 1 - в - н$. Игрок B менее уравновешен: он выигрывает партию с вероятностями $в + \epsilon$, $в$ и $в - \epsilon$ ($0 < \epsilon < \min(в, 1 - в, н, 1 - н)$) соответственно, если предыдущая партия была им выиграна, сыграна вничью или проиграна. Аналогично ведет себя вероятность его проигрыша: она равна в этих трех случаях соответственно $п - \epsilon$, $п$ и $п + \epsilon$. Очевидно, закономерности появления того или иного результата в зависимости от результата предыдущей партии у каждого из игроков могут быть описаны с помощью соответствующим образом подобранной однородной цепи Маркова с тремя возможными состояниями и с матрицами переходных вероятностей $P(A)$ (у игрока A) и $P(B)$ (у игрока B):

$$P(A) = \begin{pmatrix} в & н & п \\ в & н & п \\ в & н & п \end{pmatrix}; \quad P(B) = \begin{pmatrix} в + \epsilon & н & п - \epsilon \\ в & н & п \\ в - \epsilon & н & п + \epsilon \end{pmatrix}.$$

Заметим, что матрица $P(A)$ в действительности описывает последовательность независимых испытаний (которая, как мы видели, является частным случаем цепи Маркова), а матрица $P(B)$ иллюстрирует реаль-

ную ситуацию, подтверждающую базовую идею А. А. Маркова в построенной им простейшей цепи, в соответствии с которой вероятность определенного исхода случайного эксперимента в серии таких испытаний может зависеть от результата предыдущего исхода.

Очевидно, аналогично могут быть построены примеры марковских цепей, описывающих *динамику межотраслевой структуры трудовых ресурсов* («объект» — работник, «состояние» — отрасль экономики, в которой он занят в данный «момент времени»), *динамику социальной стратификационной структуры общества* («объект» — индивидуум, «состояние» — социальная страта, к которой он принадлежит в данный «момент времени») и т. д.

Как и всякое другое моделирование, построение моделей цепей Маркова не является самоцелью, а лишь инструментом для решения определенного класса прикладных задач. Сформулируем кратко те **основные прикладные задачи**, которые решает исследователь при анализе цепей Маркова.

Задача 1. *Прогноз распределения объектов по возможным состояниям через T тактов времени.* Другими словами, это означает, что исследователь хотел бы подсчитать по исходным данным p и P вероятности того, что анализируемый объект через заданное число тактов времени T будет находиться в заданном «состоянии» j ($j = 1, 2, \dots, N$). В примере 1 — это прогноз распределения домашних хозяйств по различным типам потребительского поведения, в примере 2 — это прогноз платежеспособности игроков через заданное число сыгранных конов, в примере 3 — это прогноз демографической структуры общества.

Задача 2. *Расчет среднего времени перехода объекта из состояния i в заданное состояние j .*

Задача 3. *Расчет вероятностей перехода из состояния i в состояние j за заданное число тактов времени t .*

Задача 4. *Расчет среднего времени пребывания объекта в заданном состоянии i .*

Задача 5. *Решение проблемы существования и вычисления стационарного распределения объектов по возможным состояниям.* Речь идет о ситуациях, когда меняющееся во времени распределение объектов по состояниям со временем стабилизируется и, начиная с некоторого «момента времени», остается практически неизменным.

5.3. Основные характеристики и свойства цепей Маркова

5.3.1. Основные характеристики

Исходными характеристиками однородной марковской цепи, как это следует из определения (5.4)–(5.6), являются:

- *вектор вероятностей начального распределения* $\mathbf{p} = (p_1, p_2, \dots, p_N)$ и
- *матрица $\mathbf{P}$ переходных вероятностей* p_{ij} ($i, j = 1, 2, \dots, N$).

Однако в процессе анализа цепи Маркова приходится сталкиваться с рядом ее *производных характеристик*, которые тем или иным способом вычисляются по исходным.

Вероятность перехода из состояния i в состояние j за t тактов времени $p_{ij}^{(t)}$ может быть подсчитана по формуле

$$p_{ij}^{(t)} = \sum_{q=1}^N p_{iq}^{(t-1)} \cdot p_{qj}, \quad (5.10)$$

которая является непосредственным следствием формулы полной вероятности (1.14), если в качестве участвующих в ней событий A и B_q рассмотреть

$$A = \{\xi_t = j \mid \xi_1 = i\} \quad \text{и} \quad B_q = \{\xi_{t-1} = q \mid \xi_1 = i\}.$$

Если по аналогии с матрицей $\mathbf{P}$ вероятностей перехода за один шаг ввести в рассмотрение матрицы $\mathbf{P}^{(q)}$ вероятностей перехода за q шагов ($q = 2, 3, \dots$), то, как легко убедиться, формула (5.10) может быть представлена в следующем матричном виде:

$$\mathbf{P}^{(t)} = \mathbf{P}^{(t-1)} \cdot \mathbf{P}. \quad (5.10')$$

Последовательно применяя эту формулу к случаям $t = 2$, $t = 3$ и т. д., получаем:

$$\begin{aligned} \mathbf{P}^{(2)} &= \mathbf{P}^{(1)} \cdot \mathbf{P} = \mathbf{P} \cdot \mathbf{P} = \mathbf{P}^2, \\ \mathbf{P}^{(3)} &= \mathbf{P}^{(2)} \cdot \mathbf{P} = \mathbf{P}^2 \cdot \mathbf{P} = \mathbf{P}^3, \\ &\dots\dots\dots \\ \mathbf{P}^{(t)} &= \mathbf{P}^t, \end{aligned} \quad (5.11)$$

т. е. для того, чтобы получить матрицу вероятностей перехода $\mathbf{P}^{(t)}$ за t шагов, надо просто возвести в степень t исходную матрицу вероятностей перехода $\mathbf{P}$.

Безусловная вероятность $p_j^{(t)}$ того, что объект в момент t находится в состоянии j , подсчитывается также с использованием формулы полной вероятности:

$$p_j^{(t)} = \sum_{i=1}^N p_i \cdot p_{ij}^{(t-1)}. \quad (5.12)$$

В формуле (5.12) вероятности $p_{ij}^{(t-1)}$, $i = 1, 2, \dots, N$, образуют j -й столбец матрицы $\mathbf{P}^{(t-1)}$, вычисляемой по формуле (5.11).

Вероятности первого возвращения объекта в заданное состояние. Введем в рассмотрение вероятность $f_{jj}^{(t)}$ первого возвращения в состояние j через t шагов, т. е.

$$f_{jj}^{(t)} = P\{\xi_{t+1} = j; \xi_t \neq j; \xi_{t-1} \neq j; \dots; \xi_2 \neq j \mid \xi_1 = j\}.$$

Опираясь на формулу сложения вероятностей, можно получить следующие рекуррентные соотношения для вычисления вероятностей $f_{jj}^{(t)}$:

$$\begin{aligned} f_{jj}^{(1)} &= p_{jj}, \\ f_{jj}^{(2)} &= p_{jj}^{(2)} - f_{jj}^{(1)} \cdot p_{jj}, \\ &\dots\dots\dots \\ f_{jj}^{(t)} &= p_{jj}^{(t)} - f_{jj}^{(1)} \cdot p_{jj}^{(t-1)} - f_{jj}^{(2)} \cdot p_{jj}^{(t-2)} - \dots - f_{jj}^{(t-1)} \cdot p_{jj}. \end{aligned} \quad (5.13)$$

Умение вычислять вероятности $f_{jj}^{(t)}$ позволяет определять и значения некоторых других характеристик, которые могут оказаться полезными при анализе марковской цепи. К таким характеристикам можно отнести, например:

- *вероятность f_{jj} того, что, выйдя из состояния j в момент времени $t = 1$, объект когда-нибудь вернется в это же состояние:*

$$f_{jj} = \sum_{t=1}^{\infty} f_{jj}^{(t)};$$

- *среднее время μ_{jj} первого возвращения в состояние j (при условии $f_{jj} = 1$):*

$$\mu_{jj} = \sum_{t=1}^{\infty} t f_{jj}^{(t)}.$$

Вероятности $f_{ij}^{(t)}$ первого достижения состояния j при выходе из состояния i ровно через t тактов времени являются естественным

обобщением вероятностей первого возвращения $f_{jj}^{(t)}$. Соотношения, позволяющие их рассчитать, также являются следствием формулы сложения вероятностей. Они имеют вид:

$$\begin{aligned} f_{ij}^{(1)} &= p_{ij}, \\ f_{ij}^{(2)} &= p_{ij}^{(2)} - f_{ij}^{(1)} \cdot p_{jj}, \\ &\dots\dots\dots \\ f_{ij}^{(t)} &= p_{ij}^{(t)} - f_{ij}^{(1)} p_{jj}^{(t-1)} - f_{ij}^{(2)} p_{jj}^{(t-2)} - \dots - f_{ij}^{(t-1)} p_{jj}. \end{aligned}$$

По этим вероятностям, в частности, могут быть подсчитаны:

- *вероятность f_{ij} того, что выйдя из состояния i в момент времени $t = 1$, объект когда-нибудь попадет в состояние j :*

$$f_{ij} = \sum_{t=1}^{\infty} f_{ij}^{(t)};$$

- *среднее время μ_{ij} первого достижения состояния j при выходе из состояния i :*

$$\mu_{ij} = \sum_{t=1}^{\infty} t f_{ij}^{(t)}.$$

5.3.2. Классификация состояний и цепей

Свойства цепи Маркова и соответственно способы решения связанных с ее анализом задач (см., например, формулировку задач 1–5 в п. 5.2) существенно зависят от того, из каких именно состояний она состоит. Поэтому приведем здесь принятую классификацию состояний конечной цепи Маркова и обусловленную ею классификацию самих марковских цепей.

Состояние j называется достижимым из состояния i , если существует такое $t_0 \geq 1$, что $p_{ij}^{(t_0)} > 0$, т. е. если имеется положительная вероятность попасть из состояния i в состояние j (включая случай $i = j$). В терминах *вероятностей первого достижения* достижимость j из i может быть определена условиями $f_{jj} > 0$ и $f_{ij} > 0$ (для $i \neq j$). Возвращаясь к нашим примерам, можно без труда убедиться в том, что в примере 5.2 все состояния достижимы из состояний $1, 2, \dots, N - 1$, а в примере 5.3 состояние j является достижимым из любого состояния i , кроме $i = 87$, только при условии $j > i$; состояние 87 — особое: из него достижимы все состояния цепи, кроме 88-го и 89-го.

Состояние i называется поглощающим, если никакое другое состояние цепи не может быть из него достигнуто, т. е. если $p_{ik} = 0$ для всех

$k \neq i$ (отсюда следует, что $p_{ii} = 1$). Мы уже имели дело с поглощающими состояниями в примерах 5.2 (состояния $i = 0$ и $i = N$), 5.3 (состояния 88 и 89) и 5.4 (состояния 6 и 7).

Свойство поглощения может быть «коллективным», а именно: **множество S состояний** называется **поглощающим** (или **замкнутым**), если никакое состояние вне S не может быть достигнуто ни из какого состояния, входящего в S . В терминах переходных вероятностей это означает, что $p_{jk} = 0$ для всех j и k таких, что j входит в S , а k не входит. Отсюда, в частности, следует, что *всякое входящее в цепь замкнутое множество состояний может быть изучено как самостоятельная (автономная) цепь Маркова независимо от прочих состояний исходной цепи*. Так, если в нашем примере 5.1 допустить существование некоего «элитного» подмножества типов потребительского поведения, за пределы которого принадлежащие этому подмножеству семьи в процессе своих переходов из одного состояния в другое *никогда не выходят*, то мы и будем иметь пример замкнутого (поглощающего) множества состояний, которое одновременно можно рассматривать как самостоятельную «подцепь» Маркова.

Возвратные состояния. Состояние i называется возвратным, если, отправляясь от него, объект с вероятностью единица когда-нибудь в него вернется, т. е. $f_{ii} = 1$ ¹. Соответственно состояние j называется **невозвратным**, если $f_{jj} < 1$; можно показать, что для этого необходимо и достаточно, чтобы $\sum_{t=1}^{\infty} p_{jj}^{(t)} < \infty$.

Периодические состояния. Состояние i называется периодическим с периодом $\tau > 1$, если $p_{ii}^{(t)} = 0$ для любого t , не кратного τ , и τ — наименьшее целое число, обладающее этим свойством (т. е. объект не может вернуться в состояние i за время, отличное от $\tau, 2\tau, 3\tau, \dots$).

Состояние принято называть эргодическим, если оно возвратное и непериодическое.

Определив различные типы состояний марковской цепи, мы переходим к **классификации самих цепей Маркова**.

Цепь называется неприводимой тогда, когда в ней нет никаких замкнутых множеств, кроме множества всех состояний. Применительно к *конечной* марковской цепи (которая и является предметом нашего

¹ В марковских цепях со счетным числом возможных состояний допускаются ситуации, когда, несмотря на возвратность состояния (и соответственно выполнение соотношения $f_{ii} = 1$), среднее время возвращения μ_{ii} оказывается бесконечно большим. В этом случае состояние i называется **возвратным нулевым**. Однако в *конечных* цепях Маркова все возвратные состояния ненулевые.

анализа в данной главе) это означает, что *неприводимая цепь Маркова должна состоять только из возвратных состояний, каждое из которых достижимо из любого другого состояния цепи*. В терминах матрицы переходных вероятностей $\mathbf{P}$ это означает, что существует такое t_0 , что *все элементы матрицы $\mathbf{P}^{t_0}$ являются положительными*. Таким образом, из определения замкнутого (поглощающего) множества состояний следует, что любое такое множество может рассматриваться как неприводимая конечная цепь Маркова.

Следует отличать определенное выше поглощающее (замкнутое) множество состояний от *множества поглощающих состояний*. Действительно, цепь Маркова может содержать одно или несколько (целое множество) поглощающих состояний. Очевидно, в этом случае она уже не может быть неприводимой хотя бы потому, что не все состояния взаимодостижимы (из поглощающего состояния ни одно другое состояние не является достижимым по определению). Ясно также, что такая цепь должна содержать невозвратные состояния, поскольку если поглощающее состояние является достижимым для какого-либо состояния j , то объект уже не может вернуться в состояние j с вероятностью единица (а значит, состояние j невозвратное). Правда, в состав сложной цепи могут входить и замкнутые множества состояний, но их, как было отмечено, можно отдельно анализировать каждое как неприводимую цепь Маркова. Поэтому представляет интерес рассмотреть класс марковских цепей, которые состоят только из поглощающих и невозвратных состояний.

Конечные марковские цепи с поглощением (или **поглощающие цепи Маркова**) и определяются как цепи, состоящие только из поглощающих и невозвратных состояний. Объекты, описываемые такой цепью Маркова, постепенно переходят из невозвратных состояний в поглощающие, находясь в невозвратных состояниях некоторое случайное время. Очевидно, если такая цепь содержит всего N состояний, из которых m -поглощающих, то при соответствующей нумерации состояний ее матрица переходных вероятностей $\mathbf{P}$ может быть представлена в виде

$$\mathbf{P} = \begin{pmatrix} \mathbf{I}_m & \vdots & \mathbf{O} \\ \dots\dots\dots & & \\ \mathbf{P}_{np} & \vdots & \mathbf{P}_{nn} \end{pmatrix}, \quad (5.20)$$

где $\mathbf{I}_m$ — единичная матрица порядка $m \times m$; $\mathbf{O}$ — матрица порядка $m \times (N - m)$, состоящая из нулей; $\mathbf{P}_{np}$ — матрица порядка $(N - m) \times m$, задающая вероятности перехода из невозвратных состояний в поглощающие; $\mathbf{P}_{nn}$ — матрица порядка $(N - m) \times (N - m)$, состоящая из вероятностей перехода между невозвратными состояниями.

Очевидно, модели зависимостей, описанные в наших примерах 5.2–5.4, относятся как раз к этому типу цепей Маркова.

Определим, наконец, для полноты картины еще один тип цепей Маркова, который, однако, в дальнейшем нам не понадобится.

Периодическая цепь — это неприводимая цепь Маркова, все возвратные состояния которой имеют один и тот же период τ . Строго говоря, в определении периодической цепи достаточно было бы потребовать периодичности любого из возвратных неприводимой цепи, т. к. по известной «теореме солидарности» (см., например, [Боровков А. А., стр. 199]) в неприводимой цепи Маркова все состояния принадлежат одному типу, и, в частности, если хоть одно из них периодически с периодом τ , то и все остальные состояния этой цепи будут иметь тот же самый период.

5.3.3. Свойства цепей Маркова

Опишем теперь ряд свойств марковских цепей, знание которых составит необходимую базу при решении прикладных задач, связанных с их анализом. Доказательства сформулированных ниже утверждений опускаются (их можно найти, например, в [Кемени Дж., Снелл Дж.] и в [Феллер В.]).

Эргодическое свойство. Пусть N — общее число состояний цепи Маркова. Для того, чтобы при любом j существовал не зависящий от i предел

$$\lim_{t \rightarrow \infty} p_{ij}^{(t)} = p_j^{(\infty)} > 0, \quad i, j = 1, 2, \dots, N,$$

необходимо и достаточно, чтобы цепь была неприводимой и непериодической, т. е. представляла бы собой одно замкнутое множество возвратных и взаимодостижимых состояний. При этом числа $p_j^{(\infty)}$ являются единственным решением системы уравнений

$$\begin{cases} p_j^{(\infty)} = \sum_{i=1}^N p_i^{(\infty)} p_{ij}, & j = 1, 2, \dots, N; \\ \sum_{j=1}^N p_j^{(\infty)} = 1. \end{cases} \quad (5.21)$$

Кроме того, существует такое h ($0 < h < 1$), что

$$|p_{ij}^{(t)} - p_j^{(\infty)}| < h^t.$$

Напомним, что $p_j^{(t)}$ — это безусловные вероятности того, что объект

в момент t находится в состоянии j (см. выше (5.12)). Так что $p_j^{(\infty)}$ есть, по-существу, вероятности попадания объекта в состояние j через большой интервал времени. При этом оказывается, что объект «забывает», откуда он начал движение, т. к. эти вероятности не зависят от начального состояния объекта. *Распределение* $(p_1^{(\infty)}, p_2^{(\infty)}, \dots, p_N^{(\infty)})$ называют **стационарным**, или **финальным**. Устойчивый режим, в который входит конечная неприводимая непериодическая цепь Маркова при $t \rightarrow \infty$, характеризуется к тому же следующими свойствами:

(а) *среднее время возвращения в состояние j*

$$\mu_{jj} = \frac{1}{p_j^{(\infty)}};$$

(б) *при больших t среднее время пребывания в состоянии j приблизительно равно $t \cdot p_j^{(\infty)}$.*

Заметим, что перенесение эргодических свойств цепи на периодический случай не представляет принципиальных трудностей, хотя и приводит к заметному усложнению формулировок.

Свойства поглощающих цепей Маркова. Эти цепи были определены в п. 5.3.2 как состоящие только из невозвратных и поглощающих состояний конечные цепи Маркова, матрица переходных вероятностей которых представима в виде (5.20). Итак, пусть из общего числа N состояний такой цепи первые m состояний отнесены к поглощающим. И пусть нас интересуют величины:

- $\alpha_{ij} = E t_{ij}$ — среднее значение времени t_{ij} пребывания в невозвратном состоянии j объекта, вышедшего из невозвратного состояния i ($i, j = m+1, m+2, \dots, N$);

- $\alpha_i = E \left(\sum_{j=m+1}^N t_{ij} \right)$ — общее время, включая время пребывания в исходном невозвратном состоянии i , которое объект проводит во всех невозвратных состояниях до попадания в какое-либо поглощающее состояние;

- γ_{ij} — вероятность попадания в поглощающее состояние j ($j = 1, 2, \dots, m$) объекта, вышедшего из невозвратного состояния i ($i = m+1, m+2, \dots, N$).

Для поглощающих цепей Маркова, задаваемых матрицами переходных вероятностей вида (5.20), имеют место следующие утверждения:

Утверждение 1. *Значения α_{ij} определяются как (i, j) -е элементы матрицы*

$$A = (I_{N-m} - P_{nn})^{-1}, \quad (5.22)$$

где I_{N-m} — единичная матрица порядка $(N-m) \times (N-m)$, а P_{nn} — матрица вероятностей перехода между невозвратными состояниями порядка $(N-m) \times (N-m)$ (см. (5.20));

Утверждение 2. Значения α_i определяются как компоненты вектора

$$\alpha = \begin{pmatrix} \alpha_1. \\ \vdots \\ \alpha_{N-m.} \end{pmatrix} = A \cdot 1, \quad (5.23)$$

где 1 — вектор-столбец порядка $(N-m) \times 1$ с компонентами, равными единице, а A — матрица, определенная соотношением (5.22);

Утверждение 3. Значения вероятностей γ_{ij} определяются как (i, j) -е элементы матрицы

$$\Gamma = A \cdot P_{nn}, \quad (5.24)$$

где A — матрица, определенная соотношением (5.22), а P_{nn} — матрица порядка $(N-m) \times m$, задающая вероятности перехода из невозвратных состояний в поглощающие (см. (5.20)).

Свойства цепи, содержащей множество невозвратных состояний J и замкнутое (поглощающее) множество возвратных состояний C . При анализе подобных цепей нас могут интересовать ответы на следующие вопросы:

- какова вероятность того, что объект, отправляясь из невозвратного состояния, все время будет находиться в множестве невозвратных состояний J ?

- как подсчитать вероятность $f_{j,C}$ того, что объект, выходя из невозвратного состояния j , когда-нибудь достигнет поглощающего множества возвратных состояний C ?

Ответы на эти вопросы дает, например, следствие теоремы 1 из [Феллер В., с. 393]:

1) В конечной цепи Маркова вероятность того, что объект все время будет находиться в множестве невозвратных состояний, равна нулю.

2) Вероятности $f_{j,C}$ того, что исходя из невозвратного состояния j объект достигнет когда-нибудь поглощающего множества C , определяются как единственное решение системы линейных уравнений

$$f_{j,C} - \sum_{k \in J} p_{jk} \cdot f_{k,C} = f_{j,C}^{(1)}, \quad j \in J. \quad (5.25)$$

В этой системе суммирование производится по всем невозвратным состояниям цепи, а $f_{j,C}^{(1)}$ — это вероятность перейти из j в любое из состояний множества C за один шаг.

5.4. Анализ некоторых задач и примеров

Теперь, когда мы проанализировали возможные типы состояний и основные свойства марковской цепи, мы можем приступить к *прикладному анализу* некоторых описанных в п. 5.2 примеров. Прикладной анализ предусматривает решение ряда конкретных задач, в том числе тех, которые были сформулированы в п. 5.2 (см. задачи 1–5).

Пример 5.1. Динамика типологии семейного потребления. Здесь нас в первую очередь будет интересовать тип и структура цепи (классификация состояний, наличие среди них замкнутых множеств возвратных состояний, невозвратные и поглощающие состояния и т. п.), а также прогноз распределения домашних хозяйств по типам потребительского поведения на заданное число тактов времени вперед (задача 1) и на отдаленное будущее (задача 5). Подавляющее большинство состояний цепи представлено *реально существующими* типами потребительского поведения (методика их статистического выявления описана в п. 12.4). Однако среди состояний может быть небольшое число *гипотетических* (*перспективных*) типов, которые в начальный (наблюдаемый) момент времени имеют «нулевое реальное представительство» (очевидно, соответствующие элементы начального распределения $\{p_i\}_{i=1,N}$ будут равны нулю). Что касается типов состояний, то из смысла задачи следует, что данная цепь, скорее всего, будет состоять из невозвратных состояний и какого-то числа поглощающих множеств и состояний. Все определится конкретными значениями переходных вероятностей p_{ij} (методика статистической оценки последних описана, например, в [Митоян А. А.]). Что касается решений задач 1 и 5, то они получаются с помощью формул — соответственно (5.11)–(5.12) и, в случае существования стационарного распределения, (5.21). Может представить практический интерес и вопрос о среднем времени пребывания домашнего хозяйства i -го типа в заданном или даже во всех невозвратных состояниях до момента его поглощения замкнутым множеством или поглощающим состоянием. Ответ на этот вопрос может быть получен с помощью соотношений (5.22) и (5.23).

Пример 5.2. Игра двух лиц до разорения. Очевидно, перед тем, как соглашаться на тот или иной регламент игры, наш участник хотел бы подсчитать *вероятность своего разорения* (при игре до разорения), а также свои *наиболее вероятные состояния* на 2-м, 3-м, 4-м и

последующих конах.

Займемся решением первой задачи. На языке цепи Маркова разорение нашего участника означает его «поглощение» состоянием $j = 0$ (оно соответствует его нулевому наличному капиталу, т. е. разорению). Несложный анализ состояний цепи показывает, что она состоит из $N - 1$ невозвратных состояний (состояния $1, 2, \dots, N - 1$) и двух поглощающих состояний (состояния 0 и N). Поглощающий тип состояний 0 и N очевиден, а невозвратность остальных состояний следует из того, что из каждого из них с положительной вероятностью достигается одно из поглощающих состояний (в этом легко убедиться, подсчитав для любого начального капитала n вероятность n -кратных последовательных проигрышей или $(N - n)$ -кратных последовательных выигрышей). Итак, наша схема относится к категории марковских цепей, *содержащих множество невозвратных состояний и замкнутое (поглощающее) множество возвратных состояний* C . В нашем случае мы имеем два таких вырожденных замкнутых множества, каждое из которых состоит из единственного поглощающего состояния. Нас интересует состояние 0 . Для определения вероятности разорения нашего участника $f_{n,0}$ составим и решим систему уравнений (5.25). Очевидно, в нашем случае

$$f_{1,0}^{(1)} = 1 - p; \quad f_{j,0}^{(1)} = 0 \quad \text{для всех } j > 1,$$

так что система (5.25) имеет вид:

$$\begin{cases} f_{1,0} - pf_{2,0} = 1 - p \\ f_{j,0} - (1 - p)f_{j-1,0} - pf_{j+1,0} = 0 \quad \text{для } j = 2, 3, \dots, N - 2 \\ f_{N-1,0} - (1 - p)f_{N-2,0} = 0 \end{cases}$$

Решение этой системы (с учетом очевидных граничных условий $f_{0,0} = 1$ и $f_{N,0} = 0$) дает:

$$f_{n,0} = \begin{cases} \left[\left(\frac{1-p}{p} \right)^N - \left(\frac{1-p}{p} \right)^n \right] / \left[\left(\frac{1-p}{p} \right)^N - 1 \right] & \text{при } p \neq \frac{1}{2}; \\ 1 - \frac{n}{N} & \text{при } p = \frac{1}{2} \end{cases}$$

(подробное решение можно найти в [Феллер В., с. 338-339]).

Пример 5.3. Демографическая модель передвижки возрастов. Рассуждения, аналогичные тем, которые мы использовали при анализе предыдущего примера, приводят нас к выводу, что данная цепь Маркова (ее матрица переходных вероятностей приведена в п. 5.2) содержит

невозвратные состояния 1–87 и два поглощающих (88-е и 89-е). Следовательно, при анализе данной модели мы можем воспользоваться описанными в п. 5.3.3 *свойствами поглощающих цепей Маркова* (5.22)–(5.24). Это позволит нам, в частности, при подходящей интерпретации состояния «пределный возраст» вычислить ряд важных демографических характеристик. Так, если удалить «пределный возраст» достаточно далеко (например, понимать под ним возраст свыше 150 лет; при этом, конечно, возрастет общее число возможных состояний), то соответствующая этому состоянию компонента вектора α , определяемого с помощью соотношения (5.23), есть не что иное, как средняя (ожидаемая) продолжительность жизни населения данной страны или региона. Если же понимать под «пределным возрастом» возраст выхода на пенсию, то аналогично можно определить среднее число лет, которое человек пребывает в пенсионном возрасте (т. е. получает пенсию), и т. д.

Пример 5.4. Модель пребывания студента в вузе. По типу цепь Маркова, описывающая эту модель, не отличается от предыдущей. Разница состоит только в том, что из каждого невозвратного состояния можно не только перейти в соседнее или в одно из двух поглощающих, но и остаться в том же самом состоянии. Рассмотрим данный пример подробнее.

Во-первых, приведем анализируемую матрицу P переходных вероятностей к виду (5.20). Для этого перенумеруем возможные состояния объекта следующим образом:

1-е состояние: «отчислен из института»;

2-е состояние: «окончил институт»;

3-е состояние: «обучение на 5-м курсе»;

4-е состояние: «обучение на 4-м курсе»;

5-е состояние: «обучение на 3-м курсе»;

6-е состояние: «обучение на 2-м курсе»;

7-е состояние: «обучение на 1-м курсе».

Зададимся для определенности численными значениями переходных вероятностей p_{ij} (данные заимствованы из [Кемени Дж., Снелл Дж.]) и

Наконец, матрица Γ попадания в поглощающие состояния

$$\Gamma = A \cdot P_{нп} = \begin{pmatrix} 1,11 & 0 & 0 & 0 & 0 \\ 0,86 & 1,11 & 0 & 0 & 0 \\ 0,67 & 0,86 & 1,11 & 0 & 0 \\ 0,52 & 0,67 & 0,86 & 1,11 & 0 \\ 0,41 & 0,52 & 0,67 & 0,86 & 1,11 \end{pmatrix} \cdot \begin{pmatrix} 0,2 & 0,7 \\ 0,2 & 0 \\ 0,2 & 0 \\ 0,2 & 0 \\ 0,2 & 0 \end{pmatrix} = \begin{pmatrix} 0,22 & 0,78 \\ 0,40 & 0,60 \\ 0,53 & 0,47 \\ 0,63 & 0,37 \\ 0,72 & 0,28 \end{pmatrix}.$$

Проанализируем полученные результаты. Компоненты вектора α свидетельствуют о том, что студенты, достигшие 5-го курса, в среднем проводят в институте еще 1,11 года (сказывается возможность остаться на 5-м курсе для повторного прохождения курса). Студенты, достигшие 4-го курса, в среднем проводят в институте еще 1,98 года (помимо упомянутого выше фактора повторного обучения в данном случае дает себя знать и перспектива досрочного отчисления). Воздействием этих двух противоположных факторов (возможность остаться «на второй год» и быть отчисленным досрочно) объясняются и средние времена пребывания в институте студентов, начинающих третий год обучения (2,65 вместо трех лет), второкурсников (3,17 вместо четырех лет) и поступивших на 1-й курс (3,58 года вместо пяти лет).

Второй столбец матрицы Γ представляет нам следующую информацию. При тех конкретных значениях переходных вероятностей, которые мы использовали в своих расчетах, лишь 28% от поступивших в вуз его успешно заканчивают. Аналогичный показатель для студентов, доучившихся до 2-го, 3-го, 4-го и 5-го курса, имеет значения соответственно 37%, 47%, 60% и 78%.

Пример 5.5. Модель игровых возможностей двух шахматистов. В этом примере главный вопрос, который нас будет интересовать, это — кто из игроков (A или B) в длительном турнире (или в серии турниров) выступит успешнее, т. е. наберет большее число очков? Для ответа на этот вопрос необходимо, по-видимому, подсчитать *стационарное распределение* по «состояниям» (выигрыш-ничья-проигрыш) для каждого из игроков. Поскольку матрица $P(A)$ переходных вероятностей, описывающая поведение игрока A , определяет *последовательность независимых испытаний* (как частный случай цепи Маркова), то и стационарное распределение по состояниям для этого игрока останется тем же самым, т. е.

$$p_1^{(\infty)}(A) = v; \quad p_2^{(\infty)}(A) = n; \quad p_3^{(\infty)}(A) = p, \quad (5.26)$$

где первое состояние соответствует выигрышу, второе — ничьей, а третье — поражению.

Система уравнений (5.21) для стационарных вероятностей игрока B будет иметь вид:

$$\begin{cases} p_1^{(\infty)}(B) \cdot (v + \varepsilon) + p_2^{(\infty)}(B) \cdot v + p_3^{(\infty)}(B) \cdot (v - \varepsilon) = p_1^{(\infty)}(B) \\ p_1^{(\infty)}(B) \cdot n + p_2^{(\infty)}(B) \cdot n + p_3^{(\infty)}(B) \cdot n = p_2^{(\infty)}(B) \\ p_1^{(\infty)}(B) + p_2^{(\infty)}(B) + p_3^{(\infty)}(B) = 1. \end{cases}$$

Решение этой системы дает:

$$p_1^{(\infty)}(B) = v + \varepsilon \frac{v - n}{1 - 2\varepsilon}; \quad p_2^{(\infty)}(B) = n; \quad p_3^{(\infty)}(B) = 1 - p_1^{(\infty)}(B) - p_2^{(\infty)}(B). \quad (5.27)$$

Сравнение соответствующих стационарных вероятностей для игроков A и B , задаваемых соотношениями (5.26) и (5.27), позволяет сделать следующий вывод: для игроков выше среднего уровня (т.е. при $v > n$) и при «не слишком большой» эмоциональности игрока B (т.е. при $\varepsilon < n$) эффективность игры шахматиста B в длительном турнире окажется выше, чем A .

* * *

Обратим внимание читателя на то, что во всех рассмотренных выше примерах переходные вероятности p_{ij} считались *априори заданными*. На практике они в большинстве случаев *оцениваются статистически*. В некоторых случаях это делается достаточно просто с помощью соответствующих относительных частот. В общем случае проблема оценки переходных вероятностей может оказаться весьма сложной (см., например, книгу Ли Ц., Джадж Д., Зельнер А. Оценивание параметров марковских моделей по агрегированным данным. М.: Статистика, 1977).

ВЫВОДЫ

1. Цепь Маркова с конечным числом состояний является простейшим обобщением *последовательности независимых испытаний*. В частности, она моделирует динамику (в дискретном времени) поведения объекта, который может находиться в одном из N состояний, в ситуациях, когда вероятность оказаться объекту в заданном состоянии j в данный момент времени t зависит только от того, в каком состоянии он находился в предыдущий момент времени $t - 1$. Если в момент времени $t - 1$ объект находился в состоянии i , то вероятность оказаться в момент времени t в состоянии j определяется так называемой *переходной вероятностью*

$p_{ij}(t)$. Цепь называется **однородной**, если эти переходные вероятности (для каждой фиксированной пары состояний i и j , $i, j = 1, 2, \dots, N$) *остаются постоянными во времени, т. е. не зависят от t* .

2. Для вероятностного описания **временной последовательности состояний** любой длины t достаточно задать (кроме матрицы $\mathbf{P} = (p_{ij})$ переходных вероятностей) распределение объектов по состояниям в первый (начальный) момент времени $\{p_1, p_2, \dots, p_N\}$, где p_i есть вероятность того, что в начальный момент времени объект находился в состоянии i .

3. **Последовательность независимых испытаний**, результатом каждого из которых может быть одно из N возможных событий, *является частным случаем однородной цепи Маркова с N возможными состояниями*, в которой матрица переходных вероятностей $\mathbf{P}$ определяется всего N различными числами, а именно, состоит из N одинаковых строк $(p_1, p_2, \dots, p_N)$.

4. Модель марковской цепи может оказаться полезной при решении ряда важных прикладных задач, среди которых:

- прогноз распределения объектов по возможным состояниям через заданное число тактов времени и в отдаленном будущем;
- расчет среднего времени и вероятности пребывания объекта в заданном состоянии или в заданном множестве состояний;
- вычисление вероятности перехода объекта из определенного состояния i в заданное множество состояний за фиксированное число тактов времени (в том числе, и за бесконечное время).

5. Способ решения этих и других задач, связанных с анализом цепей Маркова, *зависит от типа анализируемой цепи*, который, в свою очередь, определяется типологией составляющих ее состояний. К основным типам состояний марковской цепи относятся **возвратные** (вероятность когда-нибудь вернуться в возвратное состояние равна единице), **невозвратные** (вероятность когда-нибудь вернуться в невозвратное состояние меньше единицы), **поглощающие** (попав в поглощающее состояние объект с вероятностью единица остается в нем навсегда), **периодические** (объект не может вернуться в периодическое состояние за время, отличное от времени, кратного заданному периоду τ), **эргодические** (все возвратные непериодические состояния).

6. При анализе конкретной конечной цепи Маркова необходимо предварительно ее классифицировать, определив, является ли она **неприводимой** (т. е. состоящей только из возвратных взаимодостижимых состояний), и если «да», то являются ли эти состояния периодическими. В условиях неприводимой непериодической марковской цепи справедливы утверждения **эргодической теоремы**, в соответствии с которыми суще-

ствуется единственное стационарное (финальное) распределение объектов по состояниям и предлагается способ его расчета.

7. Если цепь не является неприводимой, то она может в общем случае состоять из *множества невозвратных состояний* и *какого-то числа поглощающих (замкнутых) множеств состояний* (в частном случае поглощающее множество состояний может состоять из *единственного* состояния). При решении задач, связанных с анализом подобных цепей, следует воспользоваться соотношениями (5.20), (5.22)–(5.25).

Раздел II

Основы Математической Статистики

Глава 6. Основы статистического описания и статистика нормального закона

Глава 7. Статистическое оценивание параметров

Глава 8. Статистическая проверка гипотез (статистические критерии)

ГЛАВА 6. ОСНОВЫ СТАТИСТИЧЕСКОГО ОПИСАНИЯ И СТАТИСТИКА НОРМАЛЬНОГО ЗАКОНА

6.1. Генеральная совокупность, выборка из нее и основные способы организации выборки

Закономерности, которым подчиняется исследуемая случайная величина, физически полностью обуславливаются реальным комплексом условий ее наблюдения (или эксперимента), а математически задаются соответствующим вероятностным пространством $\{\Omega, C, P\}$, или, что то же, соответствующим законом распределения вероятностей. Однако при проведении статистических исследований несколько более удобной оказывается другая терминология, связанная с понятием генеральной совокупности.

Генеральной совокупностью называют совокупность всех мыслимых наблюдений (или всех мысленно возможных объектов интересующего нас типа, с которых «снимаются» наблюдения), которые могли бы быть произведены при данном реальном комплексе условий. Поскольку в определении речь идет о всех мысленно возможных наблюдениях (или объектах), то понятие генеральной совокупности есть понятие условно-математическое, абстрактное и его не следует смешивать с реальными совокупностями, подлежащими статистическому исследованию. Так, обследовав даже все предприятия подотрасли с точки зрения регистрации значений характеризующих их технико-экономических показателей, мы можем рассматривать обследованную совокупность лишь как представителя гипотетически возможной более широкой совокупности предприятий, которые могли бы функционировать в рамках того же самого реального комплекса условий.

В практической работе элементы генеральной совокупности удобнее связывать с объектами наблюдения, чем с характеристиками этих обь-

ектов. Мы отбираем для изучения индивидуумов, семьи, предприятия, регионы, страны, но не значения их характеристик. В математической теории объекты и совокупность их характеристик не различаются и двойственность введенного определения исчезает.

Как видим, математические понятия «генеральная совокупность», «вероятностное пространство», «случайная величина» и «закон распределения вероятностей» физически полностью обуславливаются соответствующим реальным комплексом условий, а поэтому их можно считать в определенном смысле синонимами. Генеральная совокупность называется *конечной* или *бесконечной* в зависимости от того, конечна или бесконечна совокупность всех мыслимых наблюдений.

Из определения следует, что непрерывные генеральные совокупности (состоящие из наблюдений признаков непрерывной природы) всегда бесконечны. Дискретные же генеральные совокупности могут быть как бесконечными, так и конечными. Скажем, если анализируется партия из N изделий, когда каждое изделие может быть отнесено к одному из четырех сортов, исследуемой случайной величиной ξ является номер сорта случайно извлеченного из партии изделия, а множество возможных значений случайной величины состоит соответственно из четырех точек (1, 2, 3 и 4). В этом примере, очевидно, генеральная совокупность будет конечной (всего N мыслимых наблюдений). Как видим из примера, следует отличать также совокупность всех мыслимых наблюдений от *множества всех мыслимых* (или теоретически возможных) *значений* исследуемой случайной величины ξ . Наблюдений, вообще говоря, «больше», поскольку каждому фиксированному возможному значению x может соответствовать несколько или даже бесчисленное множество *мыслимых наблюдений*.

Понятие бесконечной генеральной совокупности есть математическая абстракция, как и представление о том, что измерение случайной величины можно повторить бесконечное число раз. Приблизительно бесконечную генеральную совокупность можно истолковать как предельный случай конечной, когда число объектов, порождаемых данным реальным комплексом условий, неограниченно возрастает. Так, если в только что приведенном примере вместо партии изделий рассматривать непрерывное массовое производство тех же изделий, то мы придем к понятию бесконечной генеральной совокупности. Практически же такое видоизменение равносильно требованию $N \rightarrow \infty$.

Выборка из данной генеральной совокупности — это результаты ограниченного ряда наблюдений $x_1, x_2, \dots, x_n$ случайной величины ξ . Выборку можно рассматривать как некий эмпирический аналог генеральной совокупности, то, с чем мы чаще всего на практике имеем дело, поскольку

обследование всей генеральной совокупности бывает либо слишком трудоемко (в случае больших N), либо принципиально невозможно (в случае бесконечных генеральных совокупностей).

В статистике интерпретация выборки и ее отдельных элементов допускает в зависимости от контекста два различных варианта, причем при изложении, как правило, специально не уточняется, о каком именно варианте идет речь. В целях упрощения не делается различий и в обозначениях. При первом (практическом) варианте интерпретации выборки под $x_1, x_2, \dots, x_n$ понимаются фактически наблюдаемые в данном конкретном эксперименте значения исследуемой случайной величины, т. е. конкретные числа.

В соответствии со вторым (гипотетическим) вариантом интерпретации под выборкой $x_1, x_2, \dots, x_n$ понимается последовательность случайных величин, i -й член которой (x_i) лишь обозначает результат наблюдения, который мы могли бы получить на i -м шаге n -кратного эксперимента¹, связанного с наблюдением исследуемой случайной величины ξ . Если условия эксперимента не меняются от наблюдения к наблюдению и если n -кратный эксперимент организован таким образом, что результаты наблюдения на каждом (i -м) шаге никак не зависят от предыдущих и не влияют на будущие результаты наблюдений, то, очевидно, вероятностные закономерности поведения i -го наблюдения гипотетической выборки остаются одними и теми же для всех $i = 1, 2, \dots, n$ и полностью определяются законом распределения вероятностей наблюдаемой случайной величины, т. е.

$$P\{x_i < x\} = P\{\xi < x\} = F_\xi(x). \quad (6.1)$$

При этом из взаимной независимости наблюдений выборки следует, что последовательность случайных величин

$$x_1, x_2, \dots, x_n \quad (6.2)$$

состоит из независимых компонент, т. е. их совместная функция распределения $F_{(x_1, \dots, x_n)}(z_1, \dots, z_n) = P\{x_1 < z_1, \dots, x_n < z_n\}$ может быть пред-

¹ Говоря о «шаге n -кратного эксперимента», мы не имеем в виду жесткую связь этих «шагов» с временными характеристиками проведения эксперимента. Переход от i -го наблюдения выборки к $i+1$ -му не обязательно выполняется в хронологической последовательности: выбранные для наблюдения объекты могут образовывать так называемую «пространственную выборку» и наблюдаться, например, одновременно.

ставлена в виде

$$F(x_1, \dots, x_n)(z_1, \dots, z_n) = \prod_{i=1}^n P\{x_i < z_i\} = \prod_{i=1}^n F_{\xi}(z_i). \quad (6.3)$$

Если в рамках гипотетического варианта интерпретации выборки ряд наблюдений (6.2) образует последовательность независимых и одинаково распределенных случайных величин (т. е. выполняются соотношения (6.1) и (6.3)), то выборка называется случайной.

Число n наблюдений, образующих выборку, называют **объемом выборки**.

Переход к группированным выборочным данным. Если объем выборки n велик ($n > 50$) и при этом мы имеем дело с одномерной непрерывной величиной (или с одномерной дискретной величиной, число возможных значений которой достаточно велико, скажем больше 10), то часто удобнее, с точки зрения упрощения дальнейшей статистической обработки результатов наблюдений, перейти к так называемым «группированным» выборочным данным. Этот переход осуществляется обычно следующим образом:

а) отмечают наименьшее $x_{\min}(n)$ и наибольшее $x_{\max}(n)$ значения в выборке¹;

б) весь обследованный диапазон $[x_{\min}(n), x_{\max}(n)]$ разбивается на определенное число s равных интервалов группирования; при этом количество интервалов s должно быть в пределах 7–20. Выбор количества интервалов существенно зависит от объема выборки n ; для примерной ориентации в выборе s можно пользоваться приближенной формулой $s \approx \log_2 n + 1$, которую следует воспринимать, скорее, как оценку снизу для s (особенно при больших n);

в) отмечают крайние точки каждого из интервалов $c_0, c_1, c_2, \dots, c_s$ в порядке возрастания, а также их середины $x_1^0, x_2^0, \dots, x_s^0$;

г) подсчитываются числа выборочных данных, попавших в каждый из интервалов: $\nu_1, \nu_2, \dots, \nu_s$ (очевидно, $\nu_1 + \nu_2 + \dots + \nu_s = n$); выборочные данные, попавшие на границы интервалов, либо равномерно распределяются по двум соседним интервалам, либо уславливаются относить их только к какому-либо одному из них, например, левому.

¹ Запись $x_{\min}(n)$ и $x_{\max}(n)$ подчеркивает тот факт, что наименьшее и наибольшее значения в выборке, как и все другие характеристики, построенные по выборке объема n , конечно, зависят от n . Там, где этот факт не требует напоминания, мы будем опускать n в записи выборочных характеристик.

В зависимости от конкретного содержания задачи в данную схему группирования могут быть внесены некоторые видоизменения (например, в некоторых случаях целесообразно отказаться от требования равной длины интервалов группирования; иногда крайние интервалы целесообразно делать полубесконечными слева или справа).

Репрезентативность выборки. *Сущность статистических методов состоит в том, чтобы по некоторой части генеральной совокупности (т. е. по выборке) выносить суждения о ее свойствах в целом.* Поэтому один из важнейших вопросов, от успешного решения которого зависит достоверность получаемых в результате статистического анализа выборочных данных выводов, является *вопрос репрезентативности выборки*, то есть вопрос полноты и адекватности представления ею интересующих нас свойств всей анализируемой генеральной совокупности. В практической работе одна и та же группа объектов, взятых для изучения, может рассматриваться как выборка из разных генеральных совокупностей. Так, группу семей, наудачу отобранных из панельных домов жилищно-эксплуатационной конторы (ЖЭК) одного из районов города для подробного социологического обследования, можно рассматривать и как выборку из генеральной совокупности семей (в панельных домах) данной ЖЭК, и как выборку из генеральной совокупности всех семей данного района, и как выборку из генеральной совокупности всех семей данного города, и, наконец, как выборку из генеральной совокупности всех семей города, проживающих в панельных домах. Оценка степени репрезентативности имеющихся выборочных данных существенно зависит от того, представителем какой генеральной совокупности мы рассматриваем отобранную группу семей. Ответ на этот вопрос зависит от многих факторов. В приведенном выше примере, в частности, — от наличия или отсутствия специального (быть может, скрытого) фактора, определяющего принадлежность семьи к данной ЖЭК или району в целом (таким фактором может быть, например, среднедушевой доход семьи, географическое расположение района в городе, «возраст» района и т.п.).

Основные способы организации выборки. При оценке репрезентативности выборки учитывается и то, насколько распределение в выборке существенных для изучаемого вопроса показателей характерно для анализируемой генеральной совокупности в целом. Первый путь повышения степени репрезентативности — достижение полностью случайного отбора объектов из генеральной совокупности — часто бывает труден в организационном плане. Кроме того, сочетание регулярного и случайного выбора иногда оказывается более эффективным. В любом случае способ сбора исходных данных должен тщательно планироваться и его необхо-

димо полностью описывать в отчетах о выполненной работе.

Использование для оценки репрезентативности сравнения распределений основных показателей в выборке и в генеральной совокупности также имеет свои трудности, одни из которых носят чисто статистический характер — недостаточный объем (число отобранных для обследования объектов) выборки, неразработанность методов сравнения многомерных распределений и т.п., а другие — содержательный, ведь часто заранее неизвестно, распределение каких показателей следует сравнивать при доказательстве репрезентативности.

Опишем кратко основные способы организации выборки.

Простой случайный отбор — способ извлечения n объектов из конечной генеральной совокупности N объектов, при котором каждая из возможных выборок (а всего существует, как известно, C_N^n способов отобрать n элементов из множества N элементов) имеет равную вероятность быть отобранной. На практике часто нумеруют объекты в генеральной совокупности числами от 1 до N и затем, используя таблицы случайных чисел или какой-либо другой метод, обеспечивающий равную вероятность выбора объекта (например, урну с N шарами, занумерованными цифрами от 1 до N), отбирают один за другим n объектов. Полученная таким способом выборка называется *случайной*, так как при этом обеспечивается выполнение условий случайности (6.1) и (6.3).

Простой отбор с помощью регулярной, но несущественной для изучаемого вопроса процедуры часто применяется вместо случайного отбора. На производстве изделия для контроля их качества часто отбирают в определенный период времени, что удобно с организационной точки зрения; в социологических обследованиях — по букве, с которой начинается фамилия индивидуума, проживающего в домах данной жилищно-эксплуатационной конторы, и т.п. Получаемые таким образом выборки часто называют *механическими*.

Стратифицированный (расслоенный) отбор заключается в том, что исходная генеральная совокупность объема N подразделяется на подсовокупности объема $N_1, N_2, \dots, N_R$. При этом подсовокупности не содержат общих объектов и вместе исчерпывают всю генеральную совокупность, так что $N_1 + N_2 + \dots + N_R = N$. Подсовокупности называют *стратами* или *слоями*. Когда слои определены, из каждого слоя извлекается простая случайная выборка объема соответственно $n_1, n_2, \dots, n_R$. Для того чтобы можно было полностью воспользоваться выгодами от расслоения, значения N_i/N должны быть известны. Стратифицированный отбор применяется, когда слои однородны в том смысле, что входящие в них объекты имеют близкие характеристики (средние значения которых могут

быть оценены по малым выборкам); либо когда целесообразно изучать генеральную совокупность с равной тщательностью во всех слоях; либо по организационным причинам, когда методы проведения отбора в слоях должны быть разными. Выборки, полученные таким способом, называют *стратифицированными* или *расслоенными* (иногда — *районированными*).

Частным случаем стратифицированного отбора является способ организации выборки, при котором страты (слои) генеральной совокупности выделены по *косвенному* признаку, как-то связанному с изучаемым. Так, изучая средний душевой доход семей, для получения стратифицированных выборок можно предварительно разбить исследуемую совокупность семей на группы, однородные по какой-либо из социально-экономических характеристик главы семьи (например, по заработку). В подобных случаях говорят о *типическом* способе отбора и соответственно о *типических выборках*.

Методы серийного отбора (и соответственно *серийные выборки*) используются тогда, когда удобнее назначать к обследованию не отдельные элементы генеральной совокупности, а целые «блоки» или *серии* таких элементов. Так, при проведении выборочных обследований населения способ территориально-административного деления страны и характер ведения соответствующей документации обуславливают большее удобство сплошного способа обследования целых территориальных единиц (домов, кварталов), а не отдельных семей. Подобный способ отбора часто называют также *гнездовым*.

Комбинированный (ступенчатый) отбор сочетает в себе сразу несколько из описанных выше способов отбора, образующих различные *ступени* (или фазы) выборочного обследования. Так, при выборочном обследовании условий жизни и структуры семей какого-либо города на первой ступени можно с помощью случайного отбора назначить городские районы, в которых будет производиться это обследование, затем способом механического отбора определить подлежащие обследованию жилищно-эксплуатационные конторы (ЖЭКи), а внутри ЖЭКов сделать серийную (гнездовую) выборку домов.

6.2. Основные выборочные характеристики и их свойства

В практике статистического анализа и моделирования, к сожалению, точный вид закона распределения анализируемой генеральной совокупности, как правило, бывает неизвестен. Исследователь располагает лишь выборкой из интересующей его генеральной совокупности и *вынужден*

строить все свои выводы на основании расчета ограниченного ряда выборочных характеристик. Какие же выборочные характеристики несут наибольшую информацию о генеральной совокупности, являются в этом смысле основными? К **основным выборочным характеристикам** относятся:

- выборочная (эмпирическая) функция распределения $\hat{F}^{(n)}(x)$;
- выборочная (эмпирическая) функция плотности $\hat{f}^{(n)}(x)$;
- выборочная (эмпирическая) относительная частота $\hat{p}_i^{(n)}$ появления i -го возможного значения x_i^0 дискретной случайной величины;
- выборочные начальные и центральные моменты анализируемой случайной величины ($\hat{m}_k(n)$ и $\hat{m}_k^0(n)$) и, в первую очередь, выборочное среднее значение $\bar{x}(n) = \hat{m}_1(n)$ и выборочная дисперсия $s^2(n) = \hat{m}_2^0(n)$;
- порядковые статистики $x_{(i)}(n)$ ($i = 1, 2, \dots, n$), т. е. члены ряда наблюдений выборки, расположенных в этом ряду в порядке их возрастания.

Нас будут интересовать в данном пункте определения и свойства этих выборочных характеристик. Говоря о свойствах выборочных характеристик, мы постараемся ответить в числе прочих на следующие два вопроса:

- стремится ли анализируемая выборочная характеристика к определенному пределу при неограниченном увеличении объема выборки n (т. е. при $n \rightarrow \infty$)?
- как «ведут себя» выборочные характеристики в роли случайных величин (каковыми все они являются), т. е. что можно сказать о законе распределения вероятностей каждой из них при достаточно больших (т. е. при $n \rightarrow \infty$) и ограниченных объемах выборок?

6.2.1. Выборочные (эмпирические) функции распределения, относительные частоты и функции плотности

Итак, объектом нашего анализа является генеральная совокупность, отражающая поведение случайной величины ξ с теоретической функцией распределения (вообще говоря, нам не известной) $F(x) = P\{\xi < x\}$. Мы располагаем лишь случайной выборкой (6.2) из этой генеральной совокупности.

Эмпирическим (или выборочным, т. е. построенным по выборке объема n) аналогом теоретической функции распределения $F(x)$ является

функция $\hat{F}^{(n)}(x)$, определяемая соотношением:

$$\hat{F}^{(n)}(x) = \frac{\nu(x)}{n} \quad (6.4)$$

или, в случае группированных данных (см. п. 6.1),

$$\hat{F}^{(n)}(x) = \frac{\nu_1 + \nu_2 + \dots + \nu_{i_x}}{n}, \quad (6.5)$$

где $\nu(x)$ — число наблюдаемых значений исследуемой случайной величины в выборке $x_1, x_2, \dots, x_n$, меньших x ; ν_i — число наблюдаемых значений в выборке, попавших в i -й интервал группирования, а i_x — номер самого правого из интервалов группирования, правый конец которых не превосходит x .

Из определения эмпирической функции распределения непосредственно следует объяснение часто используемого ее другого названия — «*накопленная относительная частота*».

Если анализируемая случайная величина ξ *дискретна* и имеет возможные значения $x_i^0 (i = 1, 2, \dots)$, принимаемые соответственно с вероятностями $p_i = P\{\xi = x_i^0\}$, то имеет смысл ввести понятие *выборочной* (эмпирической) *относительной частоты* $\hat{p}_i^{(n)}$, которая определяется как отношение числа $\nu_{x_i^0}$ наблюдений в выборке, в точности равных x_i^0 , к общему объему выборки, т. е.

$$\hat{p}_i^{(n)} = \frac{\nu_{x_i^0}}{n}.$$

Если ξ — *непрерывная* случайная величина (генеральная совокупность) с функцией плотности вероятности $f(x)$ и функцией распределения $F(x)$, которая всюду непрерывна и дифференцируема, то, располагая выборочными данными (выборкой) $x_1, x_2, \dots, x_n$, мы можем построить выборочный аналог функции плотности — эмпирическую (выборочную) функцию плотности $\hat{f}^{(n)}(x)$.

Для построения эмпирической (или выборочной) функции плотности $\hat{f}^{(n)}(x)$ на всей области ее определения (т. е. для всех возможных значений исследуемой величины) используют предварительно сгруппированные данные (см. п. 6.1) и полагают

$$\hat{f}^{(n)}(x) = \frac{\nu_{k(x)}}{n \cdot \Delta_{k(x)}}, \quad (6.6)$$

где $k(x)$ — порядковый номер интервала группирования, который покрывает точку x ; $\nu_{k(x)}$ — число наблюдений, попавших в этот интервал, $\Delta_{k(x)}$ — длина интервала.

Геометрическое изображение эмпирической функции плотности носит название *гистограммы*.

Пример 6.1. Анализируется выборка из ста малых предприятий региона. Цель обследования — фиксация коэффициента соотношения заемных и собственных средств (x_i) на каждом i -м предприятии. Таким образом, $n = 100$, а $i = 1, 2, \dots, n$. Результаты обследования $n = 100$ малых предприятий представлены в табл. 6.1¹.

Таблица 6.1. Коэффициенты соотношения заемных и собственных средств предприятий

5,56	5,45	5,48	5,45	5,39	5,37	5,46	5,59	5,61	5,31
5,46	5,61	5,11	5,41	5,31	5,57	5,33	5,11	5,54	5,43
5,34	5,53	5,46	5,41	5,48	5,39	5,11	5,42	5,48	5,49
5,36	5,40	5,45	5,49	5,68	5,51	5,50	5,68	5,21	5,38
5,58	5,47	5,46	5,19	5,60	5,63	5,48	5,27	5,22	5,37
5,33	5,49	5,50	5,54	5,40	5,58	5,42	5,29	5,05	5,79
5,79	5,65	5,70	5,71	5,85	5,44	5,47	5,48	5,47	5,55
5,67	5,71	5,73	4,97	5,35	5,72	5,49	5,61	5,57	5,69
5,54	5,39	5,32	5,21	5,73	5,59	5,38	5,25	5,26	5,81
5,27	5,64	5,20	5,23	5,33	5,37	5,24	5,55	5,60	5,51

Требуется построить эмпирические функции распределения $\hat{F}^{(n)}(x)$ и плотности $\hat{f}^{(n)}(x)$, а также представить их геометрическое изображение (графики).

Решение. Построим группированный ряд наблюдений:

а) определим среди наблюдений (в выборке) минимальное $x_{\min} = 5,03$ и максимальное $x_{\max} = 5,85$ значения;

б) разобьем весь диапазон $[x_{\min}, x_{\max}]$ на s равных интервалов группирования, где $s \approx \log_2 n + 1 = 3,32 \ln n + 1 = 7,62 \approx 8$, откуда ширина интервала $\Delta = \frac{x_{\max} - x_{\min}}{s} = \frac{5,85 - 5,03}{8} = 0,101$. Примем $\Delta = 0,11$;

в) определим крайние точки каждого интервала $c_0, c_1, c_2, \dots, c_s$ в порядке возрастания, а также их середины $x_1^0, x_2^0, \dots, x_s^0$.

За нижнюю границу первого интервала предлагается принять величину $c_0 = x_{\min} - \frac{\Delta}{2} = 4,97$.

¹ Результаты обследования выстроены в таблице размером 10×10 в порядке регистрации по строкам, так что 2-я строка начинается с 11-го наблюдения, 3-я — с 21-го и т. д.

Тогда $c_1 = c_0 + \Delta = 5,08$; $c_2 = c_1 + \Delta = 5,19$; ...; $c_8 = c_0 + 8\Delta = 5,85$.

Определим середины интервалов: $x_1^0 = \frac{c_0 + c_1}{2} = \frac{4,97 + 5,08}{2} = 5,03$;
 $x_2^0 = \frac{c_1 + c_2}{2} = 5,14, \dots, x_s^0 = \frac{c_{s-1} + c_s}{2} = 5,80$.

Таблица 6.2. «Группированный» ряд наблюдений

k	Интервалы $c_{k-1} - c_k$	x_k^0	ν_k	ν_k^H	$\widehat{F}^{(n)}(x)$	$\widehat{f}^{(n)}(x)$
1	4,97–5,08	5,03	2	2	0,02	0,18
2	5,08–5,19	5,14	3	5	0,05	0,27
3	5,19–5,30	5,25	12	17	0,17	1,09
4	5,30–5,41	5,36	19	36	0,36	1,73
5	5,41–5,52	5,47	29	65	0,65	2,64
6	5,52–5,63	5,58	18	83	0,83	1,64
7	5,63–5,74	5,69	13	96	0,96	1,18
8	5,74–5,85	5,80	4	100	1,00	0,36

Границы последовательных интервалов и их середины представлены в табл. 6.2.;

г) подсчитаем числа выборочных данных, попавших в каждый из интервалов: $\nu_1, \nu_2, \dots, \nu_s$, где $\nu_1 + \dots + \nu_s = n$.

Последовательно просматривая данные наблюдений, разносим значения признака по соответствующим интервалам. Так как значения могут совпадать с границами интервалов, то условимся в каждый k -й интервал включать наблюдения, **большие** или равные, чем нижняя граница интервала, и **меньшие** верхней границы, т.е. наблюдения, удовлетворяющие неравенствам $c_{k-1} \leq x < c_k$. Общее число наблюдений, отнесенных к k -му интервалу, равно частоте ν_k .

В табл. 6.2 приводятся также **накопленные частоты** ν_k^H , равные сумме частот k -го и всех предшествующих интервалов.

Для построения гистограммы на оси абсцисс откладываются крайние точки каждого из интервалов $c_0, c_1, c_2, \dots, c_s$, а по оси ординат эмпирический аналог функций плотности $\widehat{f}^{(n)}(x)$. Тогда k -му интервалу будет соответствовать прямоугольник, основанием которого является замкнутый слева интервал $[c_{k-1}, c_k)$, а высота равна $\widehat{f}^{(n)}(x) = \frac{\nu_k(x)}{\Delta \cdot n}$. Графики, соответствующие эмпирической ($\widehat{f}^{(n)}(x)$) и модельной ($\widehat{f}(x)$) плотностям, приведены на рис. 6.1. Для построения модельной кривой плотности использовалась нормальная модель закона распределения, в которую подста-

влялись вместо неизвестных параметров — среднего a и дисперсии σ^2 — значения соответствующих выборочных характеристик $\bar{x}$ и s^2 (подсчитанные в примере 6.2, см. ниже).

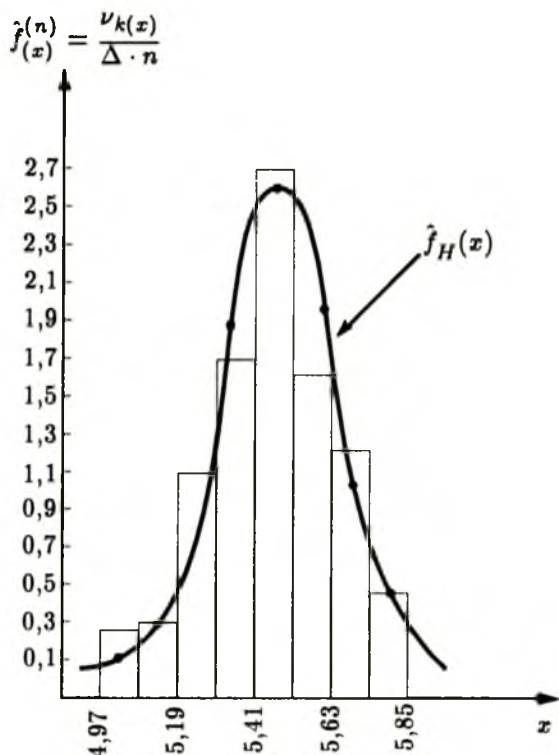


Рис. 6.1. Гистограмма

Если значения $\hat{f}_{(x)}^{(n)}(x)$ соотнести к серединам соответствующих интервалов $x_1^0, x_2^0, \dots, x_s^0$ и соединить полученные точки, то получим ломаную линию (рис. 6.2), которую называют *полигоном*.

Геометрическое представление эмпирической функции распределения $\hat{F}^{(n)}(x)$ называют *кумулянтой или кумулятивной кривой* (рис. 6.3). Для этого на оси абсцисс откладывают границы интервалов $c_0, c_1, c_2, \dots, c_s$, а по оси ординат значения $\hat{F}^{(n)}(x)$, взятые из табл. 6.2, причем значения $\hat{F}^{(n)}(x)$ относят к верхней c_k границе k -го интервала ($k = 1, 2, \dots, s$).

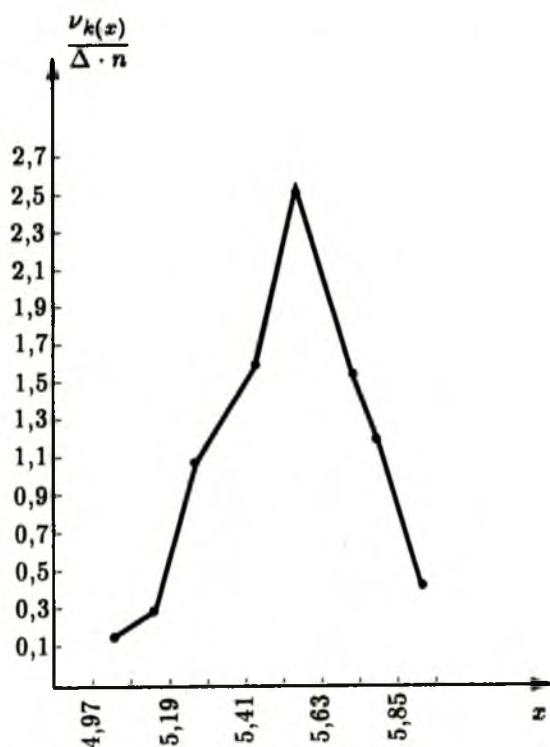


Рис. 6.2. Полигон

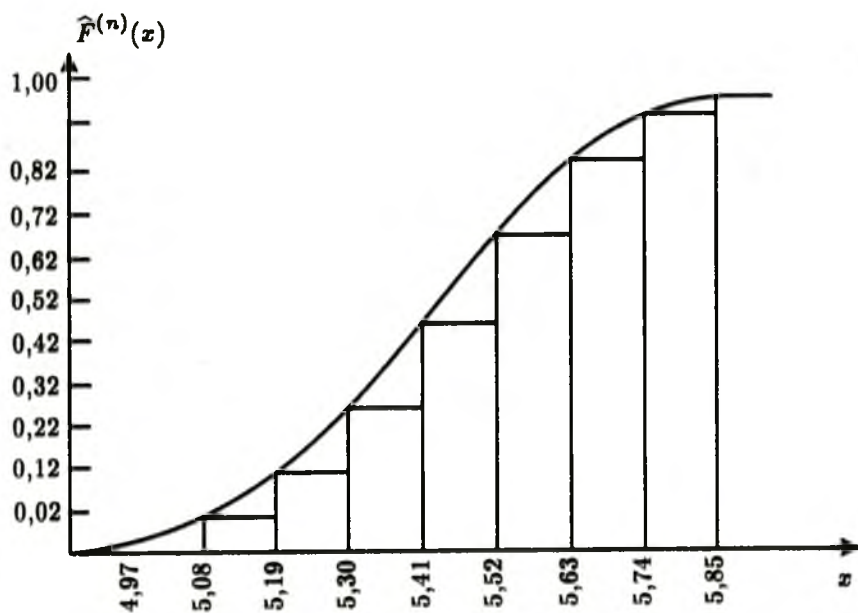


Рис. 6.3. Кумулятивная кривая

6.2.2. Выборочные аналоги начальных и центральных моментов случайной величины

В основе объяснения перехода от теоретических характеристик, которые вычисляют на базе точного знания исследуемого закона распределения, к *эмпирическим* (или *выборочным*) лежит интерпретация выборки (6.2) как модели генеральной совокупности, в которой *возможными значениями* являются наблюдаемые (т.е. практически реализованные) значения $x_1, x_2, \dots, x_n$, а в качестве вероятностей их осуществления берутся соответствующие относительные частоты их появления в выборке, т.е. величины, равные $1/n$. Таким образом, выборку можно представить в табличном виде:

x_1	x_2	$\dots$	x_n
$\frac{1}{n}$	$\frac{1}{n}$	$\dots$	$\frac{1}{n}$

Условно рассматривая ее как табличную форму задания дискретной случайной величины, возможные значения которой $x_1, x_2, \dots, x_n$ появляются с одними и теми же вероятностями $p_1 = p_2 = \dots = p_n = \frac{1}{n}$, легко представить *эмпирические аналоги* рассмотренных выше *начальных и центральных моментов* в виде:

$$\hat{m}_k = \frac{1}{n} \sum_{i=1}^n x_i^k, \quad (6.7)$$

$$\hat{m}_k^0 = \frac{1}{n} \sum (x_i - \bar{x})^k, \quad (6.8)$$

где средняя арифметическая $\bar{x}$ есть эмпирический аналог начального момента первого порядка:

$$\bar{x} = \hat{m}_1 = \frac{1}{n} \sum_{i=1}^n x_i.$$

Аналогичные формулы для *группированных выборочных данных* (см. п. 6.1) выглядят следующим образом:

$$\hat{m}_k = \frac{1}{n} \sum_{j=1}^s \nu_j (x_j^0)^k, \quad (6.7')$$

$$\hat{m}_k^0 = \frac{1}{n} \sum_{j=1}^s \nu_j (x_j^0 - \hat{m}_1)^k. \quad (6.8')$$

6.2.3. Эмпирические аналоги центра группирования генеральной совокупности

В статистической практике в качестве характеристик центра группирования значений исследуемого признака используют несколько видов средних значений, моду и медиану (см. п. 2.6.2). Рассмотрим эмпирические аналоги этих числовых характеристик, которые вычисляются по выборочным данным $x_1, x_2, \dots, x_n$.

Средняя арифметическая $\bar{x}$ (выборочная средняя) является, пожалуй, основной и наиболее употребительной характеристикой центра группирования:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (6.9)$$

Геометрическое среднее определяется по формуле:

$$\hat{G} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}. \quad (6.10)$$

Оно находит применение при расчетах темпов изменения величин и, в частности, в тех случаях, когда имеют дело с величиной, изменения которой происходят приблизительно в прямо пропорциональной зависимости с достигнутым к этому моменту уровнем самой величины (например, численность населения), или же когда имеют дело со средней из отношений, например, при расчетах «индексов цен».

Среднее гармоническое вычисляется по формуле:

$$\hat{H} = \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}}. \quad (6.11)$$

Гармоническими средними иногда пользуются в экономике при анализе средних норм времени, а также в некоторых видах индексных расчетов.

Следует отметить, что выборочное гармоническое среднее значение $\hat{H}_n$ ряда чисел $x_1, x_2, \dots, x_n$ всегда меньше выборочного геометрического среднего значения $\hat{G}_n$, которое в свою очередь меньше выборочного среднего арифметического $\bar{x}$.

Выборочная мода $\hat{x}_{\text{mod}}$ представляет собой наиболее часто встречающееся в выборочных наблюдениях значение переменной, т. е. значение, которое действительно является модным. Практическое отыскание эмпирического аналога моды по выборочным данным требует построения группированного ряда наблюдений и будет подробно рассмотрено ниже, в примере 6.2.

Медиана, точнее, ее эмпирический аналог $\hat{x}_{\text{med}}$, определяется как среднее (по местоположению) значение ранжированного, т.е. расположенного в порядке возрастания, ряда наблюдений $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ (подробнее о таких рядах см. в п. 6.3). Если n нечетно, то $\hat{x}_{\text{med}} = x_{(\frac{n+1}{2})}$, а если n четно, то $\hat{x}_{\text{med}} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$.

6.2.4. Эмпирические аналоги показателей вариации рассеивания случайной величины

Каждая из описанных ниже характеристик степени рассеивания — выборочные дисперсия, среднеквадратическое отклонение и коэффициент вариации — дает представление о том, как могут отклоняться от своего центра группирования элементы выборки.

Эмпирическую (выборочную) дисперсию s^2 можно рассматривать как приближенное значение теоретической дисперсии (см. (2.23) в п. 2.6.3):

$$s^2 = \hat{m}_2^{(0)} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (6.12)$$

Выборочное (эмпирическое) среднеквадратическое отклонение имеет вид:

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}. \quad (6.13)$$

Оно используется наряду с выборочной дисперсией s^2 для характеристики степени отклонения наблюдаемых величин $x_1, x_2, \dots, x_n$ от среднего значения $\bar{x}$ и оказывается в ряде случаев более удобным и естественным, так как s имеет ту же размерность, что и сама анализируемая случайная величина, и соответственно характеристики центра группирования.

Несколько более точными приближениями к теоретическим значениям дисперсии и среднеквадратического отклонения σ^2 и σ при *небольших объемах выборки* являются подправленные выражения $\bar{s}^2$ и $\bar{s}$ вида (подробнее о точности в определении теоретических характеристик распределения см. в п. 7.1.5):

$$\bar{s}^2 = \frac{n}{n-1} s^2; \quad \bar{s} = \sqrt{\frac{n}{n-1}} s^2. \quad (6.14)$$

Выборочный коэффициент вариации $\hat{V}$ используется в тех случаях, когда степень рассеивания естественно описывать некоторой относительной

характеристикой в соответствии со средним (см. п. 2.6.3). В частности,

$$\hat{V} = \frac{s}{\bar{x}} \cdot 100\%. \quad (6.15)$$

Как видим из (6.15), выборочный коэффициент вариации есть безразмерная величина, измеряемая в процентах.

6.2.5. Выборочные коэффициенты асимметрии и эксцесса

Выборочный коэффициент асимметрии является характеристикой степени **скошенности** (см. п. 2.6.5) и подсчитывается с помощью второго и третьего центральных выборочных моментов по формуле:

$$\hat{\beta}_1 = \frac{\hat{m}_3^{(0)}}{(\hat{m}_2^{(0)})^{\frac{3}{2}}} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{\frac{3}{2}}}. \quad (6.16)$$

Из формулы (6.16) следует, что для симметричных (относительно среднего значения $E\xi$) функций плотности распределений $\hat{\beta}_1$ должен быть близок к нулю, в то время как для распределения, гистограмма которого имеет «длинную часть», расположенную справа от ее вершины, $\hat{\beta}_1 > 0$, а если слева — то $\hat{\beta}_1 < 0$. *Выборочный эксцесс* $\hat{\beta}_2$ является, как и соответствующая теоретическая характеристика β_2 (см. п. 2.6.5), характеристикой поведения плотности (полигона) распределения в районе ее модального значения. Он подсчитывается по формуле:

$$\hat{\beta}_2 = \frac{\hat{m}_4^{(0)}}{(\hat{m}_2^{(0)})^2} - 3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s^4} - 3. \quad (6.17)$$

Напомним, что своеобразным аналогом отсчета в измерении степени островершинности служит нормальное распределение, для которого $\beta_2 = 0$. Для островершинного (по сравнению с нормальным) распределения — $\hat{\beta}_2 > 0$, а для плосковершинного — $\hat{\beta}_2 < 0$.

Пример 6.2. По данным табл. 6.2 примера 6.1 рассчитать выборочные характеристики центра группирования, рассеивания, асимметрии и эксцесса.

Решение. Для удобства вычислений целесообразно составить вспомогательную таблицу (табл. 6.3). В ней сохранены обозначения табл. 6.2 и введено одно новое: $\delta_k = x_k^0 - \bar{x}$ ($k = 1, 2, \dots, 8$).

Таблица 6.3. Вспомогательная таблица для вычисления выборочных характеристик по группированным данным

k	x_k^0	ν_k	$x_k^0 \nu_k$	$\delta_k \nu_k$	$\delta_k^2 \nu_k$	$\delta_k^3 \nu_k$	$\delta_k^4 \nu_k$	ν_k^H
1	5,03	2	10,06	-0,8712	0,37949	-0,1653	0,07201	2
2	5,14	3	15,42	-0,9768	0,31805	-0,10356	0,03372	5
3	5,25	12	63,00	-2,5872	0,55780	-0,12026	0,025928	17
4	5,36	19	101,84	-2,0064	0,21188	-0,02237	0,00236	36
5	5,47	29	158,63	0,1276	0,00056	0,00000	0,00000	65
6	5,58	18	100,44	2,0592	0,23557	0,02695	0,00308	83
7	5,69	13	73,97	2,9172	0,65462	0,14690	0,032996	96
8	5,80	4	23,20	1,3376	0,44729	0,14957	0,05002	100
Итого		100	546,56	0	2,80526	-0,08808	0,22008	

Пользуясь данными табл. 6.3, вычислим среднее арифметическое значение

$$\bar{x} = \frac{\sum_k x_k^0 \nu_k}{\sum_k \nu_k} = \frac{546,56}{100} = 5,4656.$$

Для проверки правильности вычисления $\bar{x}$ полезно убедиться в выполнении условия $\sum_k \delta_k \nu_k = \sum_k (x_k^0 - \bar{x}) \nu_k = 0$.

На основании данных табл. 6.3 найдем выборочные:

- дисперсию

$$s^2 = \frac{\sum_k (x_k^0 - \bar{x})^2 \nu_k}{\sum_k \nu_k} = \frac{\sum_k \delta_k^2 \nu_k}{\sum_k \nu_k} = \frac{2,805}{100} = 0,02805;$$

- среднеквадратическое отклонение $s = 0,167$;
- коэффициент вариации

$$\hat{V} = \frac{s}{\bar{x}} 100\% = \frac{0,167}{5,466} 100\% = 3,06\%;$$

- центральные моменты третьего и четвертого порядков:

$$\hat{m}_3^0 = \frac{\sum_k \delta_k^3 \nu_k}{\sum_k \nu_k} = \frac{-0,08808}{100} = -0,00088;$$

$$\hat{m}_4^0 = \frac{\sum_k \delta_k^4 \nu_k}{\sum_k \nu_k} = \frac{0,22008}{100} = 0,0022;$$

- коэффициент асимметрии:

$$\hat{\beta}_1 = \frac{\hat{m}_3^0}{s^3} = \frac{0,00088}{0,00468} = -0,188;$$

- коэффициент эксцесса:

$$\hat{\beta}_2 = \frac{\hat{m}_4^0}{s^4} - 3 = \frac{0,0022}{(0,028)^2} - 3 = -0,203.$$

Близость значений выборочных коэффициентов асимметрии $\hat{\beta}_1$ и эксцесса $\hat{\beta}_2$ к нулю свидетельствует в пользу выбора нормального закона распределения для анализируемой генеральной совокупности.

Как уже отмечалось, медиана $\hat{x}_{\text{med}}$ есть значение признака, приходящееся на середину ранжированного ряда наблюдений. В нашем примере $n = 100$ и $\hat{x}_{\text{med}} = \frac{x_{(50)} + x_{(51)}}{2}$. Анализ таблицы исходных данных, представленных в табл. 6.1, показывает, что наблюдения, расположенные в упорядоченном ряду на 50-м, 51-м и 52-м местах принимают одно и то же значение, равное 5,47. Другими словами, $x_{(50)} = x_{(51)} = x_{(52)} = 5,47$, следовательно, $\hat{x}_{\text{med}} = 5,47$.

Мода, как отмечалось выше, равна значению признака $\hat{x}_{\text{mod}}$, которому соответствует наибольшая частота. Как следует из табл. 6.3 наибольшая частота $\nu = 29$ соответствует значению $\hat{x}_{\text{mod}} = 5,47$.

Так как $\bar{x}$, $\hat{x}_{\text{med}}$ и $\hat{x}_{\text{mod}}$ практически не отличаются друг от друга, то есть основание предполагать, что теоретическое распределение симметрично относительно своего среднего значения, что является еще одним доводом в пользу выбора модели нормального закона.

Построение модельной кривой плотности нормального закона распределения. При расчете модельных частот ν_k^T нормального закона распределения за оценку математического ожидания a и среднеквадратического отклонения σ принимают значения соответствующих выборочных характеристик $\bar{x}$ и s , т. е. $\hat{a} = \bar{x} = 5,466$; $\hat{\sigma} = s = 0,167$.

Модельные частоты находят по формуле:

$$\nu_k^T = np_k,$$

где n — объем выборки; p_k — вероятность попадания значения нормальной случайной величины в k -й интервал.

Вероятности p_k в соответствии со свойством γ) функции распределения (см. п. 2.5.1) определяются формулой

$$p_k = P\{c_{k-1} \leq \xi < c_k\} = \Phi(c_k; \bar{x}, s^2) - \Phi(c_{k-1}; \bar{x}, s^2), \quad (6.18)$$

где $\Phi(x; a, \sigma^2)$ — функция нормального распределения, имеющего среднее значение a и дисперсию σ^2 (см. (3.8') в п. 3.1.5). Чтобы вычислять значения функции $\Phi(c_k; \bar{x}, s^2)$ для различных значений c_k и при заданных величинах $\bar{x}$ и s^2 с помощью таблиц *стандартной нормальной функции распределения* $\Phi(x) = \Phi(x; 0, 1)$ (таблицы приведены в приложении 1), воспользуемся результатом (4.9) из п. 4.4, в соответствии с которым

$$\Phi(x; a, \sigma^2) = \Phi\left(\frac{x-a}{\sigma}; 0, 1\right) = \Phi\left(\frac{x-a}{\sigma}\right). \quad (6.19)$$

Соответственно, возвращаясь к (6.18), в нашем случае имеем:

$$p_k = \Phi\left(\frac{c_k - \bar{x}}{s}\right) - \Phi\left(\frac{c_{k-1} - \bar{x}}{s}\right), \quad k = 1, 2, \dots, 8,$$

где $\bar{x}$ и s^2 — подсчитанные ранее значения выборочного среднего и выборочной дисперсии, а $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{x^2}{2}} dx$ — значение функции стандартного нормального распределения, вычисленное в точке t .

Составим табл. 6.4 для вычисления вероятностей p_k , модельных частот ν_k^T и модельной (нормальной) плотности распределения $\hat{f}_n(x)$.

График плотности модельного (нормального) закона распределения $\hat{f}_n(x)$ представлен на рис. 6.1. Этот график достаточно хорошо согласуется с гистограммой — графиком эмпирического аналога плотности $\hat{f}^{(n)}(x)$.

Таблица 6.4 Расчет модельных частот ν_k^T и модельной плотности $\hat{f}_n(x)$

$c_{k-1} - c_k$	ν_k	t_{k-1}	t_k	$\Psi(t_{k-1})$	$\Psi(t_k)$	p_k	np_k	ν_k^T	$\hat{f}_n(x)$
4,97-5,08	2	$-\infty$	- 2,39	0	0,0082	0,0082	0,8	1	0,09
5,08-5,19	3	-2,39	- 1,71	0,0082	0,0446	0,0364	3,6	4	0,36
5,19-5,30	12	-1,71	- 1,03	0,0446	0,1587	0,1141	11,4	11	1,00
5,30-5,41	19	-1,03	- 0,35	0,1587	0,3632	0,2045	20,5	21	1,91
5,41-5,52	29	- 0,35	0,33	0,3632	0,6368	0,2736	27,4	27	3,45
5,52- 5,63	18	0,33	1,02	0,6368	0,8413	0,2045	20,5	21	1,91
5,63- 5,74	13	1,02	1,70	0,8413	0,9554	0,1141	11,4	11	1,00
5,74- 5,85	4	1,70	$+\infty$	0,9554	1,0000	0,0446	4,4	4	0,36
	100	-	-	-	-	1,000	-	100	-

6.2.6. Статистическая устойчивость выборочных характеристик

Давно было замечено, что результаты отдельных наблюдений (будь то экономические, демографические, физические, метеорологические или иные наблюдения), хотя и произведенных в относительно однородных условиях, колеблются, в то время как *средние из большого числа наблюдений обнаруживают замечательную устойчивость*. К такого рода выборочным средним относятся и все введенные нами выше эмпирические (т.е. построенные по выборке) характеристики: как выборочные моменты (начальные и центральные) $\hat{m}_k$ и $\hat{m}_k^{(0)}$, так и эмпирические функция распределения $\hat{F}_\xi^{(n)}(x)$, функция плотности $\hat{f}_\xi^{(n)}(x)$ и относительные частоты $\hat{p}_i$ (при интерпретации $\hat{F}_\xi^{(n)}$, $\hat{f}_\xi^{(n)}$ и $\hat{p}_i^{(n)}$ в качестве выборочных *средних* нужно лишь помнить о возможности их выражения в терминах сумм случайных величин $\xi_1, \dots, \xi_n$, где ξ_i равно 1 и 0 в зависимости от того, попало или нет наблюдение x_i в заранее определенную нами область возможных значений (см. (4.4)–(4.6) в п. 4.2.2)).

Математическим основанием этого факта служат различные формы закона больших чисел, который позволяет теоретически обосновывать устойчивость основных эмпирических характеристик распределения — среднего значения, дисперсии, асимметрии; эксцесса, функции распределения и плотности, построенных по выборке $x_1, x_2, \dots, x_n$. При этом как всегда, когда речь идет об исследовании выборочных характеристик, мы, во-первых, подразумеваем, что имеем дело с выборкой, состоящей из *независимых* наблюдений, и, во-вторых, интерпретируем выборку во вто-

ром, *гипотетическом*, смысле как совокупность независимых наблюдений, которые *могли бы быть произведены* над анализируемой случайной величиной (см. п. 6.1, (6.1)–(6.3)). При такой интерпретации наблюдения $x_1, x_2, \dots, x_n$ суть независимые и одинаково распределенные случайные величины и к ним применимы результаты п. 4.2. Покажем, как из закона больших чисел и теоремы Я. Бернулли можно получить статистическую устойчивость основных выборочных характеристик.

а. Устойчивость выборочных начальных моментов $\hat{m}_k(n)$ и любых рациональных функций от них. Пусть существуют моменты $m_k = E\xi^k$ ($k = 1, 2, \dots, 2k_0$) всех порядков (вплоть до заданного $2k_0$) исследуемой случайной величины ξ . Тогда, применяя закон больших чисел к случайным величинам $\xi_1 = x_1^k, \xi_2 = x_2^k, \dots, \xi_n = x_n^k$, где x_i — результат i -го наблюдения исследуемого признака, мы немедленно получаем доказательство сходимости по вероятности всех выборочных начальных моментов $\hat{m}_k(n) = (\sum_{i=1}^n x_i^k)/n$ к соответствующим теоретическим моментам $m_k = E\xi^k$ ($k = 1, 2, \dots, k_0$).

Непосредственно применить закон больших чисел к *центрированным* наблюдениям $x_1 - \bar{x}(n), \dots, x_n - \bar{x}(n)$ нельзя, так как после центрирования наблюдения становятся *зависимыми*.

Однако, воспользовавшись теоремой Е. Е. Слуцкого¹ о том, что из сходимости (при $n \rightarrow \infty$) по вероятности случайных величин $\xi_i(n)$ к некоторым постоянным числам a_i ($i = 1, 2, \dots, k_0$) следует сходимость по вероятности любой рациональной функции $\varphi(\xi_1(n), \xi_2(n), \dots, \xi_{k_0}(n))$ к ее значению в точке $(a_1, a_2, \dots, a_{k_0})$, т. е. к величине $\varphi(a_1, a_2, \dots, a_{k_0})$ (если последняя существует), мы немедленно получаем доказательство сходимости по вероятности всех интересующих нас выборочных центральных моментов, асимметрии и эксцесса к соответствующим теоретическим значениям (если таковые существуют). При этом, конечно, мы учитываем, что центральные моменты, асимметрия и эксцесс являются рациональными функциями от начальных моментов (см. соотношения (2.18) в п. 2.6.1).

б. Устойчивость эмпирических функций распределения и плотности, а также относительных частот, т. е. их сходимость (при неограниченном увеличении объема выборки, по которой они построены) к соответствующим теоретическим функциям и вероятностям, следует непосредственно из (4.3')–(4.6). Продемонстрируем это на примере эмпирической функции распределения $\hat{F}^{(n)}(x)$. Введем в рассмотрение случайные

¹ Slutsky E. Über stochastische Asymptoten und Grenzwerte — Metron, 5, № 3 (1925), 3.

величины

$$\xi_i = \xi(x_i) = \begin{cases} 1, & \text{если } x_i < x; \\ 0, & \text{если } x_i \geq x. \end{cases}$$

Очевидно, $\xi_1, \xi_2, \dots, \xi_n$ — независимые, одинаково распределенные случайные величины, причем $E\xi_i = F(x)$ и $D\xi_i = F(x)(1 - F(x))$, где $F(x) = P\{\xi < x\}$ — функция распределения *исследуемой* случайной величины ξ . Легко видеть, что $\hat{F}^{(n)}(x) = [(\sum_{i=1}^n \xi_i)/n] = \bar{\xi}(n)$, и, следовательно, в соответствии с законом больших чисел $\hat{F}^{(n)}(x) \rightarrow F(x)$ (по вероятности), когда $n \rightarrow \infty$.

Таким образом, мы получили ответ на первый из вопросов, сформулированных в начале пункта и относящихся к «поведению» основных выбранных характеристик. *Тот факт, что при неограниченном увеличении объема выборки (т. е. при $n \rightarrow \infty$) все они стремятся по вероятности к соответствующим теоретическим характеристикам, дает нам основание использовать выборочные характеристики для приблизительного описания свойств всей генеральной совокупности в целом.*

6.2.7. Асимптотически-нормальный характер случайного варьирования основных выборочных характеристик

Смысл результатов предыдущего пункта заключается, в конечном счете, в том, что при осреднении большого числа (n) случайных слагаемых все менее ощущается характерный для случайных величин неконтролируемый разброс в их значениях, так что в пределе по $n \rightarrow \infty$ этот разброс как бы исчезает вовсе или, как принято говорить, *случайная величина вырождается в неслучайную*. Однако при любом *конечном* (хотя и большом) числе слагаемых n *случайный разброс у среднего арифметического этих слагаемых остается*. Поэтому и был поставлен второй вопрос в начале этого пункта относительно «поведения» основных выборочных характеристик: **что можно сказать о характере их случайного варьирования?** Ведь этот характер должен существенно зависеть от свойств той генеральной совокупности, по выборке из которой мы построили наши выборочные характеристики. Хотя последнее утверждение и верно (и, следовательно, характер случайного варьирования выборочных характеристик надо изучать отдельно для каждого типа генеральных совокупностей), однако, оказывается, что *по мере увеличения объема выборки (т. е. при $n \rightarrow \infty$) они начинают вести себя одинаковым образом независимо от специфики генеральных совокупностей, по выборкам из которых они были вычислены*.

Поэтому ответ на вопрос о характере случайного варьирования основ-

ных выборочных характеристик следует формулировать в двух вариантах в зависимости от размера объема выборки, по которой они строились:

(а) **асимптотически**, т. е. для больших объемов выборок, что математически позволяет проводить соответствующее исследование в условиях $n \rightarrow \infty$;

(б) **доасимптотически**, т. е. для ограниченных объемов выборок, что требует проведения соответствующего исследования *отдельно для каждого типа генеральной совокупности* (для нормальной, экспоненциальной, равномерной, пуассоновской и т. д.).

Ответ в условиях асимптотики (т. е. при $n \rightarrow \infty$) имеет общий характер, практически не зависит от типа анализируемой генеральной совокупности и основан на центральной предельной теореме (см. п. 4.3.1). Опираясь на различные варианты формулировки этой теоремы, можно показать, что при **больших объемах выборок все основные выборочные характеристики ведут себя как нормально распределенные случайные величины**. Конкретизация закона распределения вероятностей для *каждой конкретной выборочной характеристики* производится с помощью подсчета ее среднего значения и дисперсии. Ниже приводятся результаты для трех основных выборочных характеристик — среднего $\bar{x}$, дисперсии s^2 и функции распределения $\hat{F}^{(n)}(x)$:

$$\left\{ \begin{array}{l} \bar{x}(n) \approx N\left(a; \frac{\sigma^2}{n}\right); \\ s^2(n) \approx N\left(\left(1 - \frac{1}{n}\right)\sigma^2; \frac{2\sigma^4}{n}\left(1 + \frac{\beta_2}{2} - \frac{1 + \beta_2}{n} + \frac{\beta_2}{n^2}\right)\right); \\ \hat{F}^{(n)}(x) \approx N\left(F(x); \frac{F(x)(1 - F(x))}{n}\right). \end{array} \right. \quad (6.20)$$

В соотношениях (6.20) $N(m; \Delta^2)$, как обычно, означает нормальный закон со средним значением m и дисперсией Δ^2 , а символика « $\xi_n \approx N(m; \Delta^2)$ » означает, что случайная величина ξ_n асимптотически (по $n \rightarrow \infty$) распределена нормально с указанными величинами среднего значения (m) и дисперсии (Δ^2). Параметрами a, σ^2, β_2 и $F(x)$ обозначены соответственно среднее значение, дисперсия, коэффициент эксцесса и функция распределения анализируемой генеральной совокупности.

Вычислим приведенные в (6.20) средние значения и дисперсии выборочного среднего $\bar{x}(n)$ и эмпирической функции распределения $\hat{F}^{(n)}(x)$, а также среднее значение выборочной дисперсии $s^2(n)$ (вид ее дисперсии

приводится в (6.20) без доказательства):

$$\mathbf{E}\bar{x} = \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}x_i = \frac{1}{n} \sum_{i=1}^n a = \frac{1}{n} \cdot na = a; \quad (6.21)$$

$$\mathbf{D}\bar{x} = \mathbf{D}\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n^2} \sum_{i=1}^n \mathbf{D}x_i = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}; \quad (6.22)$$

$$\begin{aligned} \mathbf{E}s^2 &= \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right) = \frac{1}{n} \mathbf{E}\left(\sum_{i=1}^n [(x_i - a) + (a - \bar{x})]^2\right) \\ &= \frac{1}{n} \left[\sum_{i=1}^n \mathbf{E}(x_i - a)^2 + \mathbf{E}\left\{2(a - \bar{x}) \sum_{i=1}^n (x_i - a)\right\} + \sum_{i=1}^n \mathbf{E}(\bar{x} - a)^2 \right] \\ &= \frac{1}{n} \left[\sum_{i=1}^n \sigma^2 - n\mathbf{E}(\bar{x} - a)^2 \right] = \sigma^2 - \frac{1}{n} \sigma^2 = \left(1 - \frac{1}{n}\right) \sigma^2; \end{aligned} \quad (6.23)$$

$$\begin{aligned} \mathbf{E}\hat{F}^{(n)}(x) &= \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^n \xi_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}\xi_i = \frac{1}{n} \sum_{i=1}^n F(x) \\ &= \frac{1}{n} \cdot nF(x) = F(x); \end{aligned} \quad (6.24)$$

$$\begin{aligned} \mathbf{D}\hat{F}^{(n)}(x) &= \mathbf{D}\left(\frac{1}{n} \sum_{i=1}^n \xi_i\right) = \frac{1}{n^2} \sum_{i=1}^n \mathbf{D}\xi_i \\ &= \frac{1}{n^2} \cdot nF(x)(1 - F(x)) = \frac{F(x)(1 - F(x))}{n}. \end{aligned} \quad (6.25)$$

При выводе формул для $\mathbf{E}\hat{F}^{(n)}(x)$ и $\mathbf{D}\hat{F}^{(n)}(x)$ использовалось представление $\hat{F}^{(n)}(x)$ в виде $\hat{F}^{(n)}(x) = (\sum_{i=1}^n \xi_i)/n$, где

$$\xi_i = \begin{cases} 1, & \text{если } x_i < x; \\ 0, & \text{если } x_i \geq x, \end{cases}$$

и учитывалось, что $\xi_1, \xi_2, \dots, \xi_n$ — независимые, одинаково распределенные случайные величины, причем $\mathbf{E}\xi_i = 1 \cdot F(x) + 0 \cdot (1 - F(x)) = F(x)$ и $\mathbf{D}\xi_i = (1 - F(x))^2 \cdot F(x) + (0 - F(x))^2 (1 - F(x)) = F(x)(1 - F(x))$.

При выводе формулы для $\mathbf{E}s^2$ было использовано тождество:

$$\mathbf{E}\left\{2(a - \bar{x}) \sum_{i=1}^n (x_i - a)\right\} = 2\mathbf{E}\left[(a - \bar{x}) \left(\sum_{i=1}^n x_i - na\right)\right] = -2n\mathbf{E}(\bar{x} - a)^2.$$

Что касается *доасимптотического* поведения основных выборочных характеристик, то мы исследуем этот вопрос в следующем пункте этого п., но только для случая, когда выборка извлекается из нормальной генеральной совокупности.

6.2.8. Поведение выборочных характеристик в нормальной генеральной совокупности (статистика нормального закона)

Как и раньше, речь идет о свойствах выборочных характеристик, проявляющихся при повторениях выборок того же объема из одной и той же генеральной совокупности, в нашем случае — из (a, σ^2) -нормальной генеральной совокупности. Соответственно наблюдения $x_1, x_2, \dots, x_n$, образующие выборку, интерпретируются как независимые, одинаково (a, σ^2) -нормально распределенные случайные величины (см. п.6.1, соотношения (6.1)–(6.3)). Сформулируем основной результат, касающийся распределения пары основных выборочных характеристик $(\bar{x}(n), s^2(n))$ (a, σ^2) -нормальной генеральной совокупности.

Теорема Фишера. Пусть $\bar{x}(n)$ и $s^2(n)$ — соответственно выборочная средняя и выборочная дисперсия, построенные по случайной выборке $(x_1, x_2, \dots, x_n)$ из (a, σ^2) -нормальной генеральной совокупности. Тогда при *любом фиксированном объеме* выборки n (а не только при $n \rightarrow \infty$) их совместный закон распределения описывается следующим образом:

распределение $\bar{x}(n)$ подчинено

$$(a, \frac{\sigma^2}{n}) \text{ — нормальному закону;} \quad (6.26)$$

статистика $\frac{ns^2(n)}{\sigma^2}$ распределена по закону χ^2

$$\text{с } n - 1 \text{ ст. свободы;} \quad (6.27)$$

$$\bar{x}(n) \text{ и } s^2(n) \text{ статистически независимы.} \quad (6.28)$$

Мы не будем приводить здесь подробного доказательства этого важного результата. Отметим лишь, что нормальность $\bar{x}(n)$ вытекает из факта *нормальности любой линейной комбинации нормально распределенных случайных величин*, который в свою очередь получается, например, с помощью индуктивного применения формулы композиции (см. (4.13) в п.4.4) сначала к сумме *двух* нормальных случайных слагаемых, затем трех и т.д. Что касается вывода закона распределения статистики ns^2/σ^2 и статисти-

ческой независимости выборочного среднего $\bar{x}(n)$ и выборочной дисперсии $s^2(n)$, то в основе доказательства этих утверждений лежит последовательный переход сначала от наблюдений x_i к *центрированным* наблюдениям $\bar{x}_i = x_i - a$, а затем к y_i с помощью *ортogonalного* преобразования:

$$y_i = \begin{cases} \frac{1}{\sqrt{i(i+1)}} \left(\sum_{k=1}^i \bar{x}_k - i \bar{x}_{i+1} \right) & \text{при } i = 1, 2, \dots, n-1; \\ \frac{1}{\sqrt{n}} \sum_{k=1}^n \bar{x}_k & \text{при } i = n. \end{cases}$$

Несложно проверить, что полученные таким образом случайные величины y_i/σ являются независимыми, одинаково $(0,1)$ -нормально распределенными и что интересующие нас характеристики $\bar{x}(n)$ и $s^2(n)$ выразятся в терминах $y_1, y_2, \dots, y_n$ следующим образом:

$$\begin{aligned} \bar{x}(n) &= \frac{1}{\sqrt{n}} y_n + a; \\ ns^2(n) &= \sum_{i=1}^{n-1} y_i^2. \end{aligned}$$

Отсюда и будет непосредственно следовать справедливость утверждений (6.27) и (6.28).

Утверждения (6.26)–(6.28) теоремы Фишера позволяют сформулировать ряд следствий, которые окажутся весьма полезными в дальнейших наших обсуждениях, в том числе при решении задач интервального статистического оценивания (см. п. 7.5.4) и статистической проверки гипотез (гл. 8).

Следствие 1. *Среднее значение и дисперсия эмпирической дисперсии $s^2(n)$, построенной по случайной выборке объема n из (a, σ^2) -нормальной генеральной совокупности, определяются формулами:*

$$\begin{aligned} Es^2(n) &= \sigma^2 \left(1 - \frac{1}{n} \right), \\ Ds^2(n) &= \frac{2\sigma^4}{n} \left(1 - \frac{1}{n} \right). \end{aligned} \tag{6.29}$$

Выражение для $Es^2(n)$ справедливо и в *общем случае* (не только для нормальных выборок), см. выше вывод соотношения (6.23). Покажем, как результаты следствия 1 получаются из (6.27). Известно (см. п. 3.2.1), что

среднее значение и дисперсия $\chi^2(n-1)$ -распределенной случайной величины равны соответственно $n-1$ и $2(n-1)$. Поэтому

$$E\left(\frac{ns^2}{\sigma^2}\right) = n-1 \quad \text{и} \quad D\left(\frac{ns^2}{\sigma^2}\right) = 2(n-1).$$

Следовательно,

$$\frac{n}{\sigma^2} E s^2(n) = n-1 \quad \text{и} \quad \frac{n^2}{\sigma^4} D s^2(n) = 2(n-1),$$

откуда непосредственно вытекают соотношения (6.29).

Следствие 2. Если $\bar{x}(n)$ и $s^2(n)$ соответственно среднее значение и дисперсия, построенные по случайной выборке из (a, σ^2) -нормальной генеральной совокупности, то статистика

$$t = \frac{(\bar{x} - a)\sqrt{n-1}}{s} \quad \text{распределена по закону Стьюдента}$$

с $n-1$ ст.св. (6.30)

Для доказательства этого результата представим статистику t в виде

$$t = \frac{[(\bar{x} - a)/\frac{\sigma}{\sqrt{n}}]}{\sqrt{\frac{1}{n-1} ns^2/\sigma^2}}. \quad (6.31)$$

Легко устанавливается, что:

- числитель дроби (в квадратных скобках) распределен по стандартному нормальному закону (с учетом (6.26));
- подкоренное выражение в знаменателе дроби ведет себя как $\chi^2(n-1)$ -распределенная случайная величина, поделенная на число степеней свободы $n-1$ (в соответствии с (6.27));
- числитель и знаменатель дроби (6.31) статистически независимы (в соответствии с (6.28)).

Но тогда правая часть (6.31) в точности удовлетворяет определению «студентовской» случайной величины с $n-1$ степенью свободы (см. п. 3.2.2), что и завершает доказательство следствия 2.

Три следующих результата (следствия 3, 4 и 5) относятся к статистике нормального закона, связанной с анализом **двух независимых выборок** из нормальных генеральных совокупностей. А именно, пусть мы располагаем случайной выборкой

$$x_{11}, x_{12}, \dots, x_{1n_1} \quad \text{из ген. сов.} \quad N(a_1, \sigma_1^2) \quad (6.32)$$

и случайной выборкой

$$x_{21}, x_{22}, \dots, x_{2n_2} \text{ из ген. сов. } N(a_2, \sigma_2^2), \quad (6.33)$$

статистически независимой от выборки (6.32).

Нас будет интересовать поведение различных комбинаций основных выборочных характеристик $\bar{x}_1(n_1)$, $s_1^2(n_1)$, $\bar{x}_2(n_2)$ и $s_2^2(n_2)$, построенных по этим двум выборкам, где

$$\bar{x}_k(n_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ki}, \quad k = 1, 2, \quad (6.34)$$

$$s_k^2(n_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k(n_k))^2, \quad k = 1, 2. \quad (6.35)$$

Следствие 3. В условиях (6.32)–(6.34) при известных значениях дисперсий σ_1^2 и σ_2^2 имеет место следующий факт:

$$(\bar{x}_1(n_1) - \bar{x}_2(n_2)) \in N\left((a_1 - a_2); \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right). \quad (6.36)$$

В частности, при $a_1 = a_2$, $\sigma_1^2 = \sigma_2^2 = \sigma^2$ и $n_1 = n_2 = n$

$$(\bar{x}_1(n) - \bar{x}_2(n)) \in N\left(0; \frac{2\sigma^2}{n}\right). \quad (6.36')$$

Утверждение (6.36) следует из (6.26) с учетом очевидных вычислений теоретических среднего значения и дисперсии разности $(\bar{x}_1(n_1) - \bar{x}_2(n_2))$.

Следствие 4. В условиях (6.32)–(6.35) и при неизвестных, но одинаковых дисперсиях $\sigma_1^2 = \sigma_2^2 = \sigma^2$ имеет место следующий факт:

$$\frac{[(\bar{x}_1(n_1) - \bar{x}_2(n_2)) - (a_1 - a_2)] / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\sqrt{\frac{1}{n_1 + n_2 - 2} (n_1 s_1^2(n_1) + n_2 s_2^2(n_2))}} \in \text{з.р.в. Стьюдента с } n_1 + n_2 - 2 \text{ ст.св.} \quad (6.37)$$

Для того чтобы убедиться в справедливости (6.37), следует проверить правильность следующих утверждений:

а) Случайная величина

$$\gamma_1 = [(\bar{x}_1(n_1) - \bar{x}_2(n_1)) - (a_1 - a_2)] / \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

распределена по стандартному нормальному закону (следует непосредственно из (6.36) с учетом $\sigma_1^2 = \sigma_2^2 = \sigma^2$ и независимости $\bar{x}_1(n_1)$ и $\bar{x}_2(n_2)$).

б) Случайная величина

$$\gamma_2 = \left(\frac{n_1 s_1^2(n_1)}{\sigma^2} + \frac{n_2 s_2^2(n_2)}{\sigma^2} \right)$$

подчиняется з.р.в. χ^2 с $n_1 + n_2 - 2$ степенями свободы. Этот факт основан на $\chi^2(n_1 - 1)$ -распределенности $n_1 s_1^2 / \sigma^2$, $\chi^2(n_2 - 1)$ -распределенности $n_2 s_2^2 / \sigma^2$ (и то, и другое — в силу (6.27)) и независимости $s_1^2(n_1)$ и $s_2^2(n_2)$ (с учетом того, что из определения χ^2 -распределения следует, что если две χ^2 -распределенные случайные величины $\chi^2(m_1)$ и $\chi^2(m_2)$ статистически независимы, то их сумма $\chi^2(m_1) + \chi^2(m_2)$ распределена снова по закону χ^2 , но уже с $m_1 + m_2$ степенями свободы).

в) Случайная величина $\gamma_1 / \sqrt{\frac{1}{n_1 + n_2 - 2} \gamma_2}$, с одной стороны, тождественно равна левой части соотношения (6.37), а с другой стороны, распределена по закону распределения вероятностей Стьюдента с $n_1 + n_2 - 2$ степенями свободы. Тождество проверяется непосредственно, а «студентовская» распределенность этого отношения следует из статистической независимости γ_1 и γ_2 (в силу независимости выборок и (6.28)) и определения случайной величины, подчиняющейся распределению Стьюдента с m степенями свободы (см. п. 3.2.2).

Представляет самостоятельный интерес частный случай следствия 4, когда $a_1 = a_2$:

$$\frac{[\bar{x}_1(n_1) - \bar{x}_2(n_2)] / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\bar{s}} \in \text{з.р.в. Стьюдента с } n_1 + n_2 - 2 \text{ ст.св.}, \quad (6.37')$$

где величина

$$\bar{s}^2 = \frac{1}{n_1 + n_2 - 2} (n_1 s_1^2(n_1) + n_2 s_2^2(n_2))$$

является по существу оценкой неизвестной дисперсии σ^2 по объединенным наблюдениям двух выборок (6.32) и (6.33).

Результат (6.37) используют, в частности, для построения статистического критерия проверки гипотезы о равенстве средних в двух наблюдаемых нормальных генеральных совокупностях (см. гл. 8).

Следствие 5. В условиях (6.32)–(6.35) и при неизвестных, но одинаковых дисперсиях $\sigma_1^2 = \sigma_2^2 = \sigma^2$ имеет место следующий факт:

$$\frac{n_1 s_1^2(n_1)/(n_1 - 1)}{n_2 s_2^2(n_2)/(n_2 - 1)} \in \text{з.р.в. } F(n_1 - 1, n_2 - 1). \quad (6.38)$$

В правой части (6.38) обозначен закон распределения вероятностей Фишера с числом степеней свободы числителя, равным $n_1 - 1$, и числом степеней свободы знаменателя, равным $n_2 - 1$ (см. п. 3.2.3).

Справедливость (6.38) следует непосредственно из $\chi^2(n_k - 1)$ -распределенности случайных величин $n_k s_k^2(n_k)/\sigma^2$, $k = 1, 2$ (в соответствии с (6.27)) и из статистической независимости $s_1^2(n_1)$ и $s_2^2(n_2)$ (как следствие независимости выборок (6.32) и (6.33)).

6.3. Вариационный ряд и порядковые статистики

Выше отмечалось, что выборка, т.е. совокупность имеющихся у нас наблюдаемых значений $x_1, x_2, \dots, x_n$ исследуемой случайной величины ξ , является той исходной информацией, на основании которой исследователь строит свои выводы о свойствах изучаемой генеральной совокупности в целом и, в частности, составляет представление о функции распределения или плотности анализируемого закона распределения вероятностей. Оказывается, что и каждый член выборки в отдельности может доставлять важную информацию о характере анализируемого закона распределения, если наблюдения предварительно расположить в порядке возрастания. Так, например: наименьшее и наибольшее выборочные значения (соответственно $x_{\min}(n)$ и $x_{\max}(n)$) дают приближенное представление о диапазоне изменения возможных значений исследуемого признака, а их разность $(x_{\max}(n) - x_{\min}(n))$ — о степени случайного разброса его наблюдаемых значений; средний член упорядоченного ряда наблюдений — медиана $\hat{x}_{\text{med}}(n)$ — характеризует центр группирования наблюдений изучаемой случайной величины и так далее. Все это говорит о целесообразности специального рассмотрения ряда наблюдений, расположенных в порядке возрастания.

Итак, пусть $x_1, x_2, \dots, x_n$ — выборка, состоящая из n независимых наблюдений исследуемой случайной величины ξ с непрерывной функцией распределения $F_\xi(x)$ и плотностью вероятности $f_\xi(x)$. Построим новый ряд из элементов имеющейся выборки, расположив x_i ($i = 1, 2, \dots, n$) в порядке возрастания (неубывания) их значений. Обозначим член нового

ряда, стоящий на i -м месте, через $x_{(i)}$, чтобы отличать его от x_i . Тогда новый ряд будет представлен неубывающей последовательностью

$$x_{(1)}, x_{(2)}, \dots, x_{(n)}. \quad (6.39)$$

Каждый член этой последовательности ($x_{(i)}$) называется *порядковой статистикой*, а сама последовательность (6.39) — *вариационным рядом* случайной величины ξ .

Аппарат порядковых статистик широко используется как в теории и практике статистического оценивания неизвестных параметров и статистических критериев (особенно при построении устойчивых и «свободных от распределения» оценок и критериев), так и непосредственно при моделировании реальных систем и процессов. При исследовании качества оценок, критериев и моделей, полученных с использованием порядковых статистик, необходимо уметь описывать законы их распределения вероятностей в схеме *гипотетического варианта* интерпретации выборки, когда члены вариационного ряда $x_{(i)}$ интерпретируются не как конкретные числа, а как случайные величины, значения которых реализуются при повторениях выборок того же объема из той же самой генеральной совокупности. И хотя члены вариационного ряда (6.39) в отличие от членов исходной выборки уже не являются взаимно независимыми (по причине своей предварительной упорядоченности) и соответственно их частные распределения уже не являются одинаковыми, описываемыми, в частности, одной и той же плотностью $f_{\xi}(x)$ (см. (6.1)), однако, они могут быть описаны в терминах этой плотности и соответствующей функции распределения $F_{\xi}(x)$.

6.3.1. Закон распределения вероятностей i -го члена вариационного ряда

Пусть $F_{x_{(i)}}(x) = P\{x_{(i)} < x\}$ и $f_{x_{(i)}}(x)$, соответственно — непрерывные дифференцируемые функция распределения и функция плотности вероятности i -й порядковой статистики, построенной на основании случайной выборки $x_1, x_2, \dots, x_n$ из генеральной совокупности с функцией распределения $F_{\xi}(x)$ и функцией плотности вероятности $f_{\xi}(x)$. Мы хотим выразить $F_{x_{(i)}}(x)$ и $f_{x_{(i)}}(x)$ (для любого $i = 1, 2, \dots, n$) в терминах функций $F_{\xi}(x)$ и $f_{\xi}(x)$.

Реализуем следующую схему вывода требуемых соотношений. Для любого заданного значения x и некоторого достаточно малого положи-

тельного числа Δ введем в рассмотрение событие

$$A(x) = \left\{ x - \frac{\Delta}{2} \leq x_{(i)} < x + \frac{\Delta}{2} \right\}.$$

А далее мы попытаемся выразить вероятность события $A(x)$ одновременно и в терминах вероятностных характеристик случайной величины $x_{(i)}$ (что тривиально, так как $P\{A(x)\} = F_{x_{(i)}}(x + \frac{\Delta}{2}) - F_{x_{(i)}}(x - \frac{\Delta}{2})$), и в терминах *схемы n независимых наблюдений $x_1, x_2, \dots, x_n$ анализируемой случайной величины ξ* . Для этого введем в рассмотрение наряду с событием $A(x)$ еще три события:

$$A_1(x) = \left\{ \xi < x - \frac{\Delta}{2} \right\} = \left\{ x_i < x - \frac{\Delta}{2} \right\},$$

$$A_2(x) = \left\{ x - \frac{\Delta}{2} \leq \xi < x + \frac{\Delta}{2} \right\} = \left\{ x - \frac{\Delta}{2} \leq x_i < x + \frac{\Delta}{2} \right\},$$

$$A_3(x) = \left\{ \xi \geq x + \frac{\Delta}{2} \right\} = \left\{ x_i \geq x + \frac{\Delta}{2} \right\}.$$

В терминах схемы n независимых наблюдений случайной величины ξ событие $A(x)$ эквивалентно следующему событию: $i - 1$ наблюдений (из общего числа n) оказались меньше $x - \frac{\Delta}{2}$ (т. е. $i - 1$ раз состоялось событие $A_1(x)$), одно наблюдение попало в окрестность $[x - \frac{\Delta}{2}, x + \frac{\Delta}{2})$ точки x (т. е. один раз произошло событие $A_2(x)$) и остальные $n - i$ наблюдений оказались не меньшими, чем $x + \frac{\Delta}{2}$ (т. е. $n - i$ раз произошло событие $A_3(x)$). Чтобы подсчитать вероятность этого события (эквивалентного событию $A(x)$) в терминах $F_\xi(x)$ и $f_\xi(x)$, воспользуемся моделью *полиномиального распределения* (см. п. 3.1.4). Действительно по существу речь идет о производстве n независимых испытаний, в результате каждого из которых может произойти одно из событий $A_1(x)$, $A_2(x)$ или $A_3(x)$, причем вероятности этих событий по построению равны:

$$p_1 = P\{A_1(x)\} = P\{x_i < x - \frac{\Delta}{2}\} = F_\xi\left(x - \frac{\Delta}{2}\right);$$

$$p_2 = P\{A_2(x)\} = P\left\{x - \frac{\Delta}{2} \leq x_i < x + \frac{\Delta}{2}\right\} = F_\xi\left(x + \frac{\Delta}{2}\right) - F_\xi\left(x - \frac{\Delta}{2}\right);$$

$$p_3 = P\{A_3(x)\} = P\left\{x_i \geq x + \frac{\Delta}{2}\right\} = 1 - F_\xi\left(x + \frac{\Delta}{2}\right).$$

Нас интересует вероятность того, что в результате этих n независимых испытаний $i - 1$ раз произойдет событие $A_1(x)$, (т. е. $\nu_1(n) = i - 1$) один

раз — событие $A_2(x)$ (т. е. $\nu_2(n) = 1$) и $n - i$ раз — событие $A_3(x)$ (т. е. $\nu_3(n) = n - i$).

Воспользовавшись формулой (3.7) полиномиального з.р.в., имеем:

$$\begin{aligned} P\{A(x)\} &= P\{\nu_1(n) = i - 1, \nu_2(n) = 1, \nu_3(n) = n - i\} \\ &= \frac{n!}{(i-1)!1!(n-i)!} \left[F_{\xi}\left(x - \frac{\Delta}{2}\right) \right]^{i-1} \\ &\quad \times \left[F_{\xi}\left(x + \frac{\Delta}{2}\right) - F_{\xi}\left(x - \frac{\Delta}{2}\right) \right] \\ &\quad \times \left[1 - F_{\xi}\left(x + \frac{\Delta}{2}\right) \right]^{n-i}. \end{aligned} \quad (6.40)$$

С другой стороны, $P\{A(x)\}$ в терминах $F_{x(i)}(x)$ определяется в виде

$$P\{A(x)\} = F_{x(i)}\left(x + \frac{\Delta}{2}\right) - F_{x(i)}\left(x - \frac{\Delta}{2}\right). \quad (6.41)$$

Приравнивая правые части (6.40) и (6.41), деля обе части полученного таким образом выражения на Δ и устремляя Δ к нулю, получаем в пределе

$$f_{x(i)}(x) = \frac{n!}{(i-1)!(n-i)!} F_{\xi}^{i-1}(x) (1 - F_{\xi}(x))^{n-i} f_{\xi}(x), \quad i = 1, 2, \dots, n. \quad (6.42)$$

Формула (6.42) позволяет описать з.р.в. в любой порядковой статистике, зная з.р.в. анализируемой генеральной совокупности.

6.3.2. Совместные (многомерные) распределения членов вариационного ряда

Если в статистическом анализе или моделировании используются функции от *нескольких* членов вариационного ряда (например, для оценки среднего значения симметрично распределенной случайной величины может использоваться полусумма $(x_{(1)} + x_{(n)})/2$, а для оценки дисперсии — статистика, пропорциональная разности $x_{(n)} - x_{(1)}$), то для изучения свойств таких функций необходимо уметь описывать *совместные (многомерные)* з.р.в. для набора соответствующих порядковых статистик. Вычисление плотностей вероятности для таких совместных з.р.в. в терминах $f_{\xi}(x)$ и $F_{\xi}(x)$ не вызывает принципиальных трудностей, хотя и реализуется в виде весьма громоздких выражений. Мы приведем здесь

без доказательства выражение функции плотности вероятности для *любой пары порядковых статистик* $(x_{(i)}, x_{(j)})$, где $i < j$:

$$f_{(x_{(i)}, x_{(j)})}(x, y) = K(i, j, n) F_{\xi}^{i-1}(x) [F_{\xi}(y) - F_{\xi}(x)]^{j-i-1} \times [1 - F_{\xi}(y)]^{n-j} f_{\xi}(x) f_{\xi}(y), \quad (6.43)$$

где нормирующий множитель $K(i, j, n)$ определяется формулой

$$K(i, j, n) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}. \quad (6.44)$$

Продemonстрируем способ получения этой формулы на одном важном частном случае: выведем вид з.р.в. *для пары крайних членов вариационного ряда*, т. е. для $(x_{(1)}, x_{(n)})$. Для этого воспользуемся следующими очевидными соотношениями между событиями:

$$\begin{aligned} \{x_{(n)} < y\} &= \{x_{(1)} < x; x_{(n)} < y\} + \{x_{(1)} \geq x; x_{(n)} < y\} \\ &= \{x_{(1)} < x; x_{(n)} < y\} + \{x \leq x_1, x_2, \dots, x_n < y\} \\ &= \{x_{(1)} < x; x_{(n)} < y\} + \prod_{i=1}^n \{x \leq x_i < y\}. \end{aligned}$$

Соответственно имеем (учитывая независимость событий $\{x \leq x_i < y\}$ для $i = 1, 2, \dots, n$):

$$P\{x_{(n)} < y\} = P\{x_{(1)} < x, x_{(n)} < y\} + [F_{\xi}(y) - F_{\xi}(x)]^n,$$

откуда

$$F_{(x_{(1)}, x_{(n)})}(x, y) = P\{x_{(n)} < y\} - [F_{\xi}(y) - F_{\xi}(x)]^n. \quad (6.45)$$

Но событие $\{x_{(n)} < y\}$ эквивалентно произведению независимых событий $\{x_1 < y\}, \{x_2 < y\}, \dots, \{x_n < y\}$.

Поэтому из (6.45) получаем

$$F_{(x_{(1)}, x_{(n)})}(x, y) = F_{\xi}^n(y) - [F_{\xi}(y) - F_{\xi}(x)]^n \quad (6.46)$$

или, после дифференцирования обеих частей по x и по y

$$f_{(x_{(1)}, x_{(n)})}(x, y) = n(n-1)(F_{\xi}(y) - F_{\xi}(x))^{n-2} f_{\xi}(x) f_{\xi}(y). \quad (6.46')$$

Читатель может убедиться в том, что, подставляя в общую формулу (6.43)–(6.44) значения $i = 1$ и $j = n$, мы получим то же соотношение (6.46').

6.3.3. Порядковые статистики как эмпирические (выборочные) аналоги квантилей и процентных точек распределения

Из определения i -й порядковой статистики $x_{(i)}$ следует, что относительная частота (эмпирический аналог вероятности!) наблюдений выборки, меньше $x_{(i)}$, равна $(i - 1)/n$. А это означает, что порядковая статистика $x_{(i)}$ является одновременно выборочным квантилем уровня $(i - 1)/n$ и выборочной $100(1 - \frac{i-1}{n})\%$ -ной точкой анализируемого распределения.

ВЫВОДЫ

1. Понятие *генеральной совокупности* является удобным (для статистических приложений) синонимом понятий «вероятностное пространство», «случайная величина», «закон распределения вероятностей» и определяется как совокупность всех мыслимых наблюдений, которые могли бы быть сделаны в данном реальном комплексе условий.

2. *Выборка* — это статистически обследованная часть генеральной совокупности, по которой мы хотим судить об интересующих нас свойствах генеральной совокупности в целом. Вопрос представительности (репрезентативности) выборки с учетом обычной ограниченности времени и средств на ее получение требует от исследователя знания и использования различных специальных форм организации выборочных обследований.

3. Следует иметь в виду, что, говоря о выборке (или о вариационном ряде), в зависимости от контекста подразумевают *один из двух различных вариантов* интерпретации этого понятия. В первом (*практическом*) варианте под $x_1, x_2, \dots, x_n$ понимают *фактически наблюдаемые* в данном конкретном эксперименте значения исследуемой случайной величины, т. е. конкретные числа или векторы. Во втором (*гипотетическом*) варианте под $x_1, x_2, \dots, x_n$ понимают *лишь обозначение* тех n значений (чисел или векторов), *которые могли бы быть получены* при реализации n -кратного эксперимента (наблюдения) в реальном комплексе условий, индуцирующем исследуемую генеральную совокупность. В последнем случае сами x_i и любые функции от них выступают уже не в качестве конкретных чисел или векторов, а в качестве *случайных величин*, поскольку их значения варьируют неконтролируемым образом при каждом новом извлечении выборки того же объема n из той же самой генеральной совокупности.

4. В практике статистического анализа и моделирования точный вид з.р.в. анализируемой случайной величины, а соответственно и точные значения ее основных теоретических числовых характеристик (среднего, дисперсии и т.п.) бывают неизвестны. Исследователь располагает лишь

выборкой из интересующей его генеральной совокупности, а потому *вынужден строить все свои выводы на основании ограниченного ряда выборочных характеристик*. К основным выборочным характеристикам одномерной случайной величины относятся: эмпирические функции распределения и плотности ($\hat{F}^{(n)}(x)$, $\hat{f}^{(n)}(x)$), выборочная относительная частота $\hat{p}_i^{(n)}$ появления i -го возможного значения, выборочное среднее значение ($\bar{x}(n)$), выборочная дисперсия ($s^2(n)$), члены вариационного ряда ($x_{(i)}$).

Б. Осмысленное и обоснованное использование в статистическом анализе основных выборочных характеристик требует знания их свойств, в том числе ответов на вопросы: 1) стремятся ли они к значениям своих теоретических аналогов при неограниченном увеличении объема выборки (т.е. при $n \rightarrow \infty$)? 2) что можно сказать о з.р.в. каждой из выборочных характеристик? Оказывается, в *асимптотическом* смысле (т.е. при $n \rightarrow \infty$) ответы на эти вопросы практически не зависят от специфики генеральной совокупности, из которой извлекалась выборка, и являются следующими: *основные выборочные характеристики с ростом объема выборки стремятся к своим теоретическим аналогам и ведут себя при этом как нормально распределенные случайные величины*.

В. Исследование поведения основных выборочных характеристик при *конечных* (ограниченных) объемах выборок требует учета специфики генеральной совокупности, из которой извлечена выборка. Пример результатов такого исследования дает *теорема Фишера*, описывающая поведение пары основных выборочных характеристик ($\bar{x}(n)$, $s^2(n)$) для случая нормальной генеральной совокупности. Ряд полезных следствий этой теоремы (*статистика нормального закона*) используется в дальнейшем при построении интервальных оценок и статистических критериев проверки гипотез.

7. Полезным приемом в исследовании свойств анализируемой случайной величины является расположение имеющихся наблюдений $x_1, x_2, \dots, x_n$ в порядке их возрастания: $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, так что $\{x_{(i)}\}_{i=1, \bar{n}}$ — это те же самые наблюдения $\{x_i\}_{i=1, \bar{n}}$, только *упорядоченные* (т.е. $x_{(i)} \leq x_{(i+1)}$ для всех $i = 1, 2, \dots, n-1$). Ряд $\{x_{(i)}\}_{i=1, \bar{n}}$ называют *вариационным рядом*, а его члены — *порядковыми статистиками*. Они также относятся к основным выборочным характеристикам и, в частности, широко используются при построении так называемых *непараметрических оценок и непараметрических критериев*.

ГЛАВА 7. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ

Одна из главных целей, которые ставит перед собой исследователь, приступая к статистической обработке исходных данных, заключается в том, чтобы *добиться удобной лаконичности* в описании интересующих его свойств исследуемой совокупности (или исследуемого явления), т. е. представления множества обрабатываемых данных в виде сравнительно небольшого числа *сводных характеристик*, построенных на основании этих исходных данных. При этом желательно, чтобы потеря информации, существенной для принятия решения, была минимальной. Упомянутые сводные характеристики являются функциями от исходных результатов наблюдения $x_1, x_2, \dots, x_n$ и называются *статистиками* (таким образом, мы уже имеем, как минимум, три различных обиходных варианта термина «статистика»: научная дисциплина, исходная информация и любая функция от результатов наблюдения). В предыдущей главе мы уже имели дело с примерами таких сводных характеристик (статистик): к ним относятся все выборочные (*эмпирические*) характеристики генеральной совокупности — средние значения, дисперсии, коэффициенты эксцесса и асимметрии, наконец, эмпирическая функция распределения и эмпирическая плотность.

Добиться лаконичности в описании информации, содержащейся в массиве обрабатываемых данных, помогает целый *набор прикладных методов* математической статистики: выбор и обоснование математической модели механизма изучаемого явления; изучение свойств анализируемой системы или механизма функционирования с помощью небольшого числа сводных выборочных характеристик (среднего, дисперсии и т. п.); наглядное представление (визуализация) исходных данных с целью формирования рабочих гипотез о механизме изучаемого явления; анализ относительных частот, выборочных функций распределения и плотности и другие методы *описательной* (или дескриптивной) *статистики*; анализ природы обрабатываемых данных; описание интересующих исследователя ста-

статистических связей между анализируемыми признаками и т.д. Все эти методы в той или иной степени опираются на *две основные составные части математического аппарата статистики*: 1) *теорию статистического оценивания неизвестных значений параметров*, участвующих в описании анализируемой модели; 2) *теорию проверки статистических гипотез о параметрах или природе анализируемой модели*.

Изложению основных элементов первой из этих двух составных частей аппарата математической статистики и посвящена настоящая глава.

7.1. Начальные сведения о задаче статистического оценивания параметров

7.1.1. Постановка задачи

Пусть мы располагаем исходными статистическими данными — выборкой

$$x_1, x_2, \dots, x_n \quad (7.1)$$

из исследуемой генеральной совокупности и пусть интересующие нас свойства этой генеральной совокупности могут быть описаны с помощью уравнения (*математической модели*)

$$M(x, \Theta) = 0, \quad (7.2)$$

где x_i — i -е наблюдение в выборке (7.1), x — текущее (т.е. подставляемое по нашему усмотрению) значение исследуемого случайного признака, $\Theta = (\theta^{(1)}, \dots, \theta^{(k)})$ — k -мерный параметр, участвующий в записи модели (7.2), значения которого нам были не известны до получения выборки (7.1).

Задача статистического оценивания неизвестных параметров Θ по выборке заключается, грубо говоря, в построении такой k -мерной векторной функции $\hat{\Theta}(x_1, \dots, x_n) = (\hat{\theta}^{(1)}(x_1, \dots, x_n), \dots, \hat{\theta}^{(k)}(x_1, \dots, x_n))^T$ от имеющихся у нас наблюдений (7.1), которая давала бы в *определенном смысле наиболее точные приближенные значения* для истинных (не известных нам) значений параметров $\Theta = (\theta^{(1)}, \dots, \theta^{(k)})^T$. Здесь не уточняется пока, в каком именно смысле приближенные значения $\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(k)}$ соответственно параметров $\theta^{(1)}, \dots, \theta^{(k)}$ являются наилучшими.

В качестве моделей (7.2) могут, например, рассматриваться модели законов распределения вероятностей.

Пример 7.1. Пусть нашей целью является исследование закона распределения наблюдаемой непрерывной одномерной случайной величины ξ с неизвестной плотностью вероятности $f_{\xi}(x)$ и пусть предварительный анализ природы исходных данных (7.1) привел нас к выводу, что этот закон может быть описан *нормальной моделью*. В этом случае

$$\Theta = (\theta^{(1)}, \theta^{(2)})^T, \quad \text{где } \theta^{(1)} = a = E\xi \quad \text{и} \quad \theta^{(2)} = \sigma^2 = D\xi;$$

$$M(x; \Theta) = f_{\xi}(x) = \frac{1}{\sqrt{2\pi\theta^{(2)}}} e^{-\frac{(x-\theta^{(1)})^2}{2\theta^{(2)}}}$$

и можно показать, что вся информация о параметрах $\theta^{(1)}$ и $\theta^{(2)}$ (а следовательно, и о всей модели) содержится всего в двух статистиках — $\bar{x}$ и s^2 , где

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (7.3)$$

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (7.4)$$

Выше показано, что есть основания использовать статистики (7.3) и (7.4) в качестве приближенных значений (оценок) параметров a и σ^2 , поскольку по мере роста объема выборки, т. е. при $n \rightarrow \infty$, эти оценки сходятся по вероятности к соответствующим истинным значениям a и σ^2 . Однако вопрос о том, являются ли эти оценки *наилучшими*, пока остается открытым. Чтобы иметь возможность обсуждать этот вопрос, нам придется ввести и обсудить ряд понятий.

7.1.2. Статистики, статистические оценки, их основные свойства

Любая функция $\gamma(x_1, \dots, x_n)$ от результатов наблюдения $x_1, x_2, \dots, x_n$ исследуемой случайной величины ξ называется *статистикой*. Мы уже познакомились с целым рядом статистик: выборочное среднее; выборочная ковариационная матрица; выборочные коэффициенты асимметрии $\hat{\beta}_1$ и эксцесса $\hat{\beta}_2$; эмпирические функции распределения $\hat{F}^{(n)}(x)$ и плотности $\hat{f}^{(n)}(x)$.

Статистика $\hat{\Theta}$, используемая в качестве приближенного значения неизвестного параметра Θ , называется *статистической оценкой*. Так, например, статистики $\bar{x}, s^2, \hat{\beta}_1$ и $\hat{\beta}_2$ можно рассматривать как статистические оценки соответственно параметров $a = E\xi, D\xi, \beta_1$ и β_2 , поскольку,

как это следует из п. 7.2 все эти статистики при $n \rightarrow \infty$ сходятся по вероятности к истинным значениям соответствующих параметров.

Обращаем внимание читателя на тот факт, что, говоря о статистиках и статистических оценках, мы используем всегда *гипотетический* вариант интерпретации выборки, т. е. вариант, при котором под $x_1, \dots, x_n$ подразумеваются лишь обозначения тех n значений исследуемого признака ξ , которые мы *могли бы получить*, проводя n -кратный случайный эксперимент (или производя n независимых наблюдений) в данном реальном комплексе условий. Следовательно, *все статистики и статистические оценки являются случайными величинами*: при переходе от одной выборки к другой (даже в рамках одной и той же генеральной совокупности) конкретные значения статистической оценки, подсчитанные по одной и той же формуле (7.3) (т. е. значения, полученные с помощью подстановки в эту формулу соответственно различных конкретных значений аргумента), будут подвержены некоторому неконтролируемому разбросу. Правда, значения статистической оценки, подсчитанные по разным выборкам, хотя и подвержены случайному разбросу, но должны (если наша оценка «хороша») концентрироваться около истинного значения оцениваемого параметра.

Возникает вопрос о требованиях, которые следует предъявить к статистическим оценкам, чтобы эти оценки были в каком-то определенном смысле надежными. Эти требования формулируются обычно с помощью следующих трех свойств оценок: *состоятельности, несмещенности и эффективности*.

7.1.3. Состоятельность

Оценка $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$ неизвестного параметра θ называется *состоятельной*, если по мере роста числа наблюдений n (т. е. при $n \rightarrow \infty$) она стремится по вероятности к оцениваемому значению θ , т. е. если для любого сколь угодно малого $\varepsilon > 0$ $P\{|\hat{\theta} - \theta| > \varepsilon\} \rightarrow 0$ при $n \rightarrow \infty$ (если оцениваемый параметр θ векторный, то для состоятельности соответствующей векторной оценки $\hat{\theta}$ требуется состоятельность отдельно всех ее компонент). Заметим, что достаточными условиями состоятельности оценки $\hat{\theta}_n$ параметра θ являются следующие ее свойства:

1) смещение $B_n = E\hat{\theta}_n - \theta$ оценки $\hat{\theta}_n$ равно нулю ($B_n = 0$) или стремится к нулю при $n \rightarrow \infty$, т. е. $\lim_{n \rightarrow \infty} B_n = 0$;

2) дисперсия оценки $D\hat{\theta}_n$ удовлетворяет условию $\lim_{n \rightarrow \infty} D\hat{\theta}_n = 0$.

Докажем, что эти условия являются достаточными. Согласно нера-

венству Чебышева, можно записать

$$P(|\hat{\theta}_n - E\hat{\theta}_n| < \varepsilon) \geq 1 - \frac{D\hat{\theta}_n}{\varepsilon^2}.$$

Подставляя выражение $E\hat{\theta}_n = B_n - \theta$, получим

$$P(|\hat{\theta}_n - B_n - \theta| < \varepsilon) \geq 1 - \frac{D\hat{\theta}_n}{\varepsilon^2}.$$

Переходя к пределу при $n \rightarrow \infty$ с учетом того, что по условию $B_n = 0$ или $\lim_{n \rightarrow \infty} B_n = 0$, получим

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1,$$

что и требовалось доказать.

Все упомянутые выше оценки ($\bar{x}$, s^2 , $\hat{F}^{(n)}(x)$, $\hat{f}^{(n)}(x)$ и т.д.) являются, как это показано в п. 6.2, состоятельными оценками соответствующих параметров.

С одной стороны, требование состоятельности представляется необходимым для того, чтобы оценка имела практический смысл (так как в противном случае увеличение объема исходной информации не будет «приближать нас к истине»), а потому это свойство должно проверяться в первую очередь.

С другой стороны, свойство состоятельности — это *асимптотическое* (по числу наблюдений n) свойство, т. е. оно может проявляться лишь при столь больших объемах выборок, до которых мы в нашей практике не «добираемся». Кроме того, в большинстве ситуаций можно предложить несколько состоятельных оценок одного и того же параметра. Например, оценки $\hat{\theta}_1 = \bar{x}$ и $\hat{\theta}_2 = (x_{\min} + x_{\max})/2$ являются состоятельными оценками неизвестного истинного среднего значения $\theta = E\xi$ симметрично распределенной случайной величины, если это среднее θ существует (здесь $\bar{x}$ — выборочное среднее, определенное по формуле (7.3), а $x_{\min}$ и $x_{\max}$ — соответственно минимальное и максимальное значения среди n наблюдений исследуемого признака).

Более того, можно привести примеры состоятельных оценок, *уводящих сколь угодно далеко от истины*. Представим себе, что произведена достаточно большая случайная выборка работников некоторой отрасли экономики с регистрацией величины их заработной платы, т. е. мы располагаем выборкой вида (7.1). Заказчик поручает статистику построить оценку теоретического значения средней заработной платы $a = E\xi$ анализируемой совокупности с единственным требованием, чтобы эта оценка

была состоятельной. И представим себе, что статистик по тем или иным причинам *заинтересован в завышении реального результата* (или в этом заинтересован заказчик, о чем он проинформировал статистика). Тогда статистик может предложить в качестве состоятельной следующую оценку:

$$\hat{a}_n = \begin{cases} a_0 & \text{при } n < N - 1; \\ \bar{x} & \text{при } n \geq N - 1, \end{cases}$$

где a_0 — любое удобное статистику (или заказчику) значение, $\bar{x}$ — выборочное среднее значение, построенное по выборке (7.1), а N — общее число работников в анализируемой отрасли. Как легко видеть, состоятельность оценки $\hat{a}_n$ следует из состоятельности $\bar{x}$ (так как в процессе $n \rightarrow \infty$ мы перейдем рубеж $n = N - 1$ и, следовательно, асимптотические свойства $\hat{a}_n$ будут определяться асимптотическими свойствами $\bar{x}$). В то же время при всех *реальных* значениях объема выборки n оценка $\hat{a}_n$ будет давать *заданное* (удобное заказчику или статистику) значение a_0 , которое никак не связано со свойствами анализируемой генеральной совокупности. И чтобы разоблачить недобросовестность статистика (заказчика), потребуется в данном случае произвести сплошное обследование платы работников анализируемой области.

Все это говорит о том, что свойства состоятельности недостаточно для полной характеристики надежности оценки. Поэтому его надо дополнить рассмотрением двух других свойств.

7.1.4. Несмещенность

Оценка $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$ неизвестного параметра θ называется *несмещенной*, если при любом объеме выборки n результат ее осреднения по всем возможным выборкам данного объема приводит к точному истинному значению оцениваемого параметра, т. е. $E\hat{\theta} = \theta$ (если оцениваемый параметр Θ векторный, то для несмещенности соответствующей векторной оценки Θ требуется несмещенность отдельно всех ее компонент).

Проверим, например, являются ли оценки (7.3) и (7.4) несмещенными оценками параметров $a = E\xi$ и $\sigma^2 = D\xi$:

$$E\bar{x} = E\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n} \sum_{i=1}^n Ex_i = \frac{1}{n} \sum_{i=1}^n a = a;$$

$$\begin{aligned}
Es^2 &= \frac{1}{n} E \left[\sum_{i=1}^n ((x_i - a) - (\bar{x} - a))^2 \right] \\
&= \frac{1}{n} E \left[\sum_{i=1}^n (x_i - a)^2 - 2(\bar{x} - a) \sum_{i=1}^n (x_i - a) + \sum_{i=1}^n (\bar{x} - a)^2 \right] \\
&= \frac{1}{n} E \left[\sum_{i=1}^n (x_i - a)^2 - 2(\bar{x} - a)(n\bar{x} - na) + n(\bar{x} - a)^2 \right] \\
&= \frac{1}{n} E \left[\sum_{i=1}^n (x_i - a)^2 - n(\bar{x} - a)^2 \right] \\
&= \frac{1}{n} \left[\sum E(x_i - a)^2 - n \cdot E(\bar{x} - a)^2 \right] \\
&= \frac{1}{n} \left(n\sigma^2 - n \cdot \frac{\sigma^2}{n} \right) = \sigma^2 \left(1 - \frac{1}{n} \right).
\end{aligned}$$

В ходе вычисления Es^2 мы воспользовались тем фактом, что в случае независимых и одинаково распределенных, с дисперсией σ^2 , наблюдений $x_1, x_2, \dots, x_n$ имеем

$$D\left(\frac{x_1 + \dots + x_n}{n}\right) = \frac{1}{n^2} \sum_{i=1}^n Dx_i = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Мы видим, что $\bar{x}$ (6.21) является несмещенной оценкой параметра a , в то время как оценка s^2 (6.23) параметра σ^2 имеет отрицательное смещение, равное σ^2/n .

В отличие от состоятельности несмещенность характеризует «доасимптотические» свойства оценки, т.е. является характеристикой ее хороших свойств при каждом *конечном* объеме выборки. Удовлетворение требованию несмещенности устраняет систематическую погрешность оценивания, которая, вообще говоря, зависит от объема выборки n и в случае состоятельности оценки стремится к нулю при $n \rightarrow \infty$. Если смещение оценки удалось определить, то оно легко устраняется. Так, в нашем примере для устранения смещения достаточно перейти к оценке

$$\frac{s^2}{n-1} = \frac{n}{n-1} s^2,$$

которая, как легко понять, уже будет несмещенной. Из сказанного следует также, что требование несмещенности (при соблюдении требования

состоятельности) особенно существенно при малом количестве наблюдений.

7.1.5. Эффективность

Представим себе, что мы имеем две состоятельные и несмещенные векторные оценки $\hat{\Theta}_1(x_1, \dots, x_n)$ и $\hat{\Theta}_2(x_1, \dots, x_n)$ неизвестного векторного параметра Θ . Для возможности геометрической интерпретации примера будем полагать размерность k векторного параметра равной двум ($k = 2$). Для анализа свойств двух конкурирующих оценок будем производить многократное (в данном примере двадцатикратное) оценивание неизвестного параметра $\Theta = (\theta^{(1)}, \theta^{(2)})^T$ каждым из двух рассматриваемых способов. С этой целью подсчитываем значения $\hat{\Theta}_{1i}$ и $\hat{\Theta}_{2i}$ ($i = 1, 2, \dots, 20$), являющиеся результатом подстановки в функции $\hat{\Theta}_1$ и $\hat{\Theta}_2$ i -й по порядку выборки объема n , т. е. извлекаем первую выборку объема n — $x_{11}, x_{12}, \dots, x_{1n}$, вставляем эти наблюдения в качестве аргументов функций $\hat{\Theta}_1$ и $\hat{\Theta}_2$, получаем первую пару оценок $\hat{\Theta}_{11}$ и $\hat{\Theta}_{21}$; затем извлекаем вторую выборку объема n — $x_{21}, x_{22}, \dots, x_{2n}$, вставляем эти наблюдения в качестве аргументов тех же функций $\hat{\Theta}_1$ и $\hat{\Theta}_2$ — получаем вторую пару оценок $\hat{\Theta}_{12}$ и $\hat{\Theta}_{22}$, и т.д. На рис. 7.1. по горизонтальной оси отложены первая компонента неизвестного (оцениваемого) параметра ($\theta^{(1)}$) и первые компоненты ее двух оценок ($\hat{\theta}_{1i}^{(1)}$ на рис. 7.1. а и $\hat{\theta}_{2i}^{(1)}$ на рис. 7.1. б), а по вертикальной оси — вторая компонента неизвестного (оцениваемого) параметра ($\theta^{(2)}$) и вторые компоненты ее двух оценок ($\hat{\theta}_{1i}^{(2)}$ на рис. 7.1. а и $\hat{\theta}_{2i}^{(2)}$ на рис. 7.1. б).

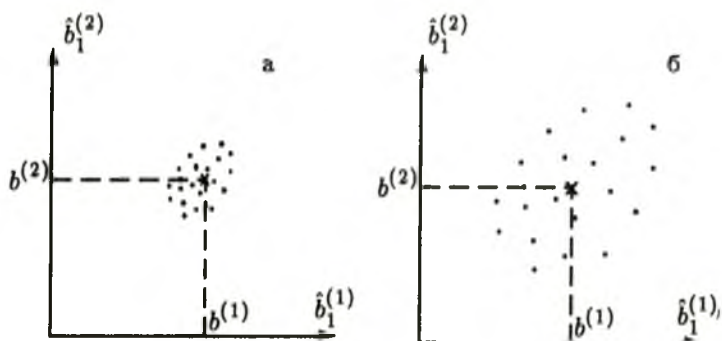


Рис. 7.1. Два способа состоятельного несмещенного оценивания многомерного параметра $\Theta = (\theta^{(1)}, \theta^{(2)})$, характеризующегося разной эффективностью: а) более эффективная оценка; б) менее эффективная оценка

Таким образом, взаимное расположение точки $(\hat{\theta}_{1i}^{(1)}, \hat{\theta}_{1i}^{(2)})$ и крестика $(\theta^{(1)}, \theta^{(2)})$ на рис. 7.1, а дает наглядное представление о близости оценки $\hat{\theta}_{1i}$, полученной первым способом с использованием i -й выборки, к истинному значению оцениваемого параметра Θ (аналогичная картина для второго способа оценивания представлена на рис. 7.1, б). Более тесная концентрация оценок, полученных первым способом, около истинного значения, очевидно, склонит нас к мысли о большей эффективности оценки $\hat{\theta}_1$ по сравнению с оценкой $\hat{\theta}_2$.

Именно этот критерий как мера разброса оцененных значений $\hat{\theta}_2$ около истинного значения Θ в соответствующем k -мерном пространстве и положен в основу определения эффективности оценки. *Оценка $\hat{\theta}$ параметра Θ называется эффективной, если она среди всех прочих оценок того же самого параметра обладает наименьшей мерой случайного разброса относительно истинного значения оцениваемого параметра.* Эффективность является решающим свойством, определяющим качество оценки, и оно, вообще говоря, не предполагает обязательного соблюдения свойства несмещенности.

Остается уточнить, как именно измеряется степень случайного разброса значений оценки $\hat{\theta}$ относительно истинной величины параметра Θ .

В случае, когда Θ — *скаляр* (т.е. размерность оценки $k = 1$), в качестве такой естественной меры берется средний квадрат отклонения, т.е. величина $E(\hat{\Theta} - \Theta)^2$, что для несмещенных оценок совпадает с их дисперсией, так как в этом случае $D\hat{\Theta} = E(\hat{\Theta} - E\hat{\Theta})^2 = E(\hat{\Theta} - \Theta)^2$.

В случае, когда оценка $\hat{\Theta}$ — *вектор* (т.е. размерность оценки $k \geq 2$), в качестве меры отклонения от истинного значения векторного параметра Θ обычно рассматривается ковариационная матрица оценки $\hat{\Theta}$, т.е. симметричная и неотрицательно-определенная матрица размера $k \times k$, которую мы будем обозначать $\Sigma(\hat{\Theta})$. Соответственно оценка $\hat{\theta}_1$ параметра Θ считается более эффективной, чем оценка $\hat{\theta}_2$, если существуют из ковариационные матрицы $\Sigma(\hat{\theta}_1)$ и $\Sigma(\hat{\theta}_2)$ и матрица $\Delta\Sigma = \Sigma(\hat{\theta}_1) - \Sigma(\hat{\theta}_2)$ является неотрицательно-определенной.

Для векторных оценок возможны случаи, когда, несмотря на существование матриц $\Sigma(\hat{\theta}_1)$ и $\Sigma(\hat{\theta}_2)$, нельзя ответить на вопрос, какая из двух оценок эффективнее в указанном выше смысле. Эта неопределенность устраняется, если в качестве меры отклонения векторной несмещенной оценки $\hat{\theta}$ от истинного значения оцениваемого параметра Θ рассматривать не саму ковариационную матрицу оценки $\Sigma(\hat{\theta})$, а ее определитель $\det \Sigma(\hat{\theta})$ (*обобщенная дисперсия*) или след $\text{tr} \Sigma(\hat{\theta})$.

Подчеркнем тот факт, что именно эффективность оценки, измеряемая средним квадратом ее отклонения от истинного значения пара-

метра, является решающим свойством, определяющим ее качество, ее надежность. Бывают ситуации, когда требования несмещенности и эффективности оценки оказываются несовместимыми. И тогда, как правило, следует руководствоваться критерием эффективности.

Приведем такой пример. Речь идет о ситуации, в которой смещенная оценка оказывается лучше несмещенной в смысле среднего квадрата ошибки.

Пусть мы располагаем выборкой (7.1) из $(a; \sigma^2)$ -нормальной генеральной совокупности (см. пример 7.1) и нас интересует наилучшая (в смысле среднего квадрата ошибки) оценка параметра $\sigma^2 = D\xi$. Мы уже знаем (см. п. 7.1.4), что оценка

$$\bar{s}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

является несмещенной оценкой теоретической дисперсии σ^2 . Воспользовавшись соотношением (6.29), получаем:

$$\begin{aligned} E(\bar{s}^2 - \sigma^2)^2 &= D\bar{s}^2 = D\left(\frac{n}{n-1}s^2\right) = \frac{n^2}{(n-1)^2} \cdot Ds^2 = \frac{n^2}{(n-1)^2} \cdot \frac{2(n-1)}{n^2} \sigma^4 \\ &= \frac{2\sigma^4}{n-1}. \end{aligned}$$

Рассмотрим в качестве альтернативы к оценке $\bar{s}^2$ класс оценок $\{\alpha\bar{s}^2\}$, где α — любое положительное число, и попробуем подобрать α_0 так, чтобы $E(\alpha_0\bar{s}^2 - \sigma^2)^2 = \min_{\alpha}$.

Рассмотрим выражение среднего квадрата ошибки для оценок этого класса:

$$\begin{aligned} E(\alpha\bar{s}^2 - \sigma^2)^2 &= E[\alpha(\bar{s}^2 - \sigma^2) - \sigma^2(1 - \alpha)]^2 \\ &= \alpha^2 \frac{2\sigma^4}{n-1} + (1 - \alpha)^2 \sigma^4. \end{aligned}$$

Легко устанавливается, что правая часть этого выражения, рассматриваемая (с точностью до умножения на положительное число σ^4) как квадратный трехчлен относительно α :

$$\left(\frac{2}{n-1} + 1\right)\alpha^2 - 2\alpha + 1$$

имеет минимум в точке $\alpha = \frac{n-1}{n+1}$. Подставляя это значение α , мы имеем

оценку

$$s^{*2} = \frac{n-1}{n+1} \bar{s}^2 = \frac{1}{n+1} \sum_{i=1}^n (x_i - \bar{x})^2$$

и ее средний квадрат ошибки

$$E(s^{*2} - \sigma^2)^2 = \left(\frac{n-1}{n+1}\right)^2 \frac{2\sigma^4}{n-1} + \left(1 - \frac{n-1}{n+1}\right)^2 \sigma^4 = \frac{2}{(n+1)} \sigma^4.$$

Легко видеть, что величина $\frac{2}{(n+1)} \sigma^4$ будет всегда меньше величины $E(\bar{s}^2 - \sigma^2)^2 = \frac{2\sigma^4}{n-1}$, а это и означает, что *смещенная оценка s^{*2} параметра σ^2 оказалась точнее несмещенной оценки $\bar{s}^2$* .

7.2. Функция правдоподобия. Количество информации, содержащейся в n независимых наблюдениях относительно неизвестного значения параметра

Пусть (7.1) — выборка, состоящая из n независимых одномерных наблюдений, извлеченная из исследуемой генеральной совокупности. Закон распределения вероятностей наблюдаемой случайной величины ξ описывается функцией $f(x, \Theta)$, зависящей от неизвестного параметра Θ , причем мы будем понимать под $f(x, \Theta)$ вероятность $P\{\xi = x\}$, если ξ дискретная, и значение плотности вероятности в точке x , если ξ непрерывна. Если рассматривать выборку (7.1) в гипотетическом смысле, то каждая конкретная выборка $(x_1^*, x_2^*, \dots, x_n^*)$ представляется определенной точкой в n -мерном пространстве выборок переменных $x_1, x_2, \dots, x_n$ и имеет смысл говорить о совместном распределении вектора $X = (x_1, \dots, x_n)$. Поскольку при гипотетическом варианте понимания случайной выборки $x_1, x_2, \dots, x_n$ суть независимые и одинаково распределенные случайные величины, то для любого заданного набора значений $x_1^*, x_2^*, \dots, x_n^*$ их совместная плотность (вероятность) будет

$$L(x_1^*, x_2^*, \dots, x_n^*; \theta) = f(x_1^*; \theta) \cdot f(x_2^*; \theta) \cdot \dots \cdot f(x_n^*; \theta). \quad (7.5)$$

Таким образом, функция $L(X^*, \Theta)$, определенная равенством (7.5), задает вероятность получения, при извлечении выборки объема n именно наблюдений $x_1^*, \dots, x_n^*$ (или величину, пропорциональную вероятности получения выборочных значений в непосредственной близости от точки X^* в

непрерывном случае). Поэтому, чем больше значение $L(X^*, \Theta)$, тем *правдоподобнее* (или более вероятна) система наблюдений $X^* = (x_1^*, \dots, x_n^*)$ при заданном значении параметра Θ . Отсюда и название функции L — *функция правдоподобия*.

Функция правдоподобия в зависимости от постановки задач и целей исследования может рассматриваться либо как функция параметра Θ (при заданных фиксированных наблюдениях $x_1^*, \dots, x_n^*$), либо как функция текущих значений наблюдений $x_1, \dots, x_n$ (при заданном фиксированном значении параметра Θ), либо как функция обеих переменных X и Θ .

Интересно попытаться проследить характер изменения вероятности (7.5) в зависимости от изменения значения параметра Θ . Очевидно, чем резче проявляется эта зависимость, тем больше информации заключено в конкретных значениях величин X и Θ друг о друге. При этом под информацией о неизвестном параметре Θ , содержащейся в случайной величине X , понимают степень уменьшения неопределенности, касающейся неизвестного значения Θ , после наблюдения над данной случайной величиной. Если по наблюденному значению X^* случайной величины X можно с вероятностью единица точно восстановить значение параметра Θ , то это значит, что случайная величина (или ее наблюдение) содержит максимально возможную информацию о параметре. И наоборот, если распределение (7.5) случайной величины X одно и то же при всех значениях параметра Θ , то нет никаких оснований делать какие-либо заключения о Θ по результатам наблюдений этой случайной величины (ситуация нулевой информации относительно значения неизвестного параметра, содержащейся в наблюдении). Чувствительность случайной величины к параметру может быть измерена величиной изменения распределения этой случайной величины при изменении значения параметра. Наиболее часто используемой характеристикой, на основании которой измеряют изменение в распределении (7.5) при небольшом изменении значения параметра Θ , является так называемое *количество информации Фишера* (содержащееся в наблюдениях $X = (x_1, x_2, \dots, x_n)$), которое определяется для скалярного параметра θ (т.е. при размерности параметра Θ , равной единице) следующим образом:

$$I(\theta; X) = E \left[\left(\frac{d \ln L}{d\theta} \right)^2 \right] = \int \left(\frac{d}{d\theta} \ln L(X; \theta) \right)^2 L(X; \theta) dX. \quad (7.6)$$

Интеграл в правой части (7.6) означает на самом деле *n-кратное* интегрирование подинтегрального выражения по всем возможным значениям $x_1, x_2, \dots, x_n$; соответственно, под dX понимается *n-мерный «элемент интегрирования»*, т.е. $dX = dx_1 dx_2 \dots dx_n$.

Учитывая независимость и одинаковую распределенность наблюдений $x_1, x_2, \dots, x_n$, получаем

$$I(\theta; X) = n \int \left(\frac{d \ln f(x; \theta)}{d\theta} \right)^2 f(x; \theta) dx = n I(\theta; x). \quad (7.6')$$

Если параметр $\Theta = (\theta^{(1)}, \dots, \theta^{(k)})$ k -мерный, причем $k \geq 2$, то вместо количества информации (7.6) рассматривается *информационная матрица Фишера* $I(\Theta, X)$ размерности $k \times k$ с элементами

$$I_{ij}(\Theta, X) = E \left[\left(\frac{\partial \ln L}{\partial \theta^{(i)}} \times \frac{\partial \ln L}{\partial \theta^{(j)}} \right) \right]. \quad (7.7)$$

Эти понятия были введены Фишером в 20-х годах нашего столетия.

Воспользовавшись формулой (7.6), нетрудно подсчитать количество информации $I(\theta; x)$, содержащееся в одном наблюдении о параметре θ , в ряде конкретных примеров.

1. Одномерная величина $X = x$ подчинена (a, σ^2) -нормальному закону с плотностью $\varphi(x; a, \sigma^2)$, в котором среднее значение $a = \theta$ — неизвестный параметр, а дисперсия известна. Тогда

$$\begin{aligned} I(a; x) &= \int_{-\infty}^{\infty} \left(\frac{d \ln \varphi(x; a, \sigma^2)}{da} \right)^2 \varphi(x; a, \sigma^2) dx \\ &= \int_{-\infty}^{\infty} \left(\frac{x-a}{\sigma^2} \right)^2 \varphi(x; a, \sigma^2) dx = \frac{1}{\sigma^2}. \end{aligned} \quad (7.8)$$

Результат, естественно, интерпретируется следующим образом: чем больше дисперсия σ^2 , тем больше разброс в наблюдаемых значениях исследуемой случайной величины, тем меньше информации о величине ее среднего значения заключено в одном наблюдении.

2. Одномерная случайная величина $X = x$ подчинена (a, σ^2) -нормальному закону с плотностью $\varphi(x; a, \sigma^2)$, в котором среднее значение a известно, а дисперсия σ^2 является неизвестным параметром. Тогда

$$\begin{aligned} I(\sigma^2; x) &= \int_{-\infty}^{\infty} \left(\frac{d \ln \varphi(x; a, \sigma^2)}{d\sigma^2} \right)^2 \varphi(x; a, \sigma^2) dx \\ &= \int_{-\infty}^{\infty} \left[\frac{(x-a)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right]^2 \varphi(x; a, \sigma^2) dx = \frac{1}{2\sigma^4}. \end{aligned} \quad (7.9)$$

3. Одномерная случайная величина $X = x$ подчинена гамма-распределению с параметрами (a, b) , причем параметр a известен, а b является неизвестным параметром. Тогда

$$I(b; x) = \int_0^{\infty} \left(\frac{d \ln f_{\gamma(a,b)}(x)}{db} \right)^2 \times f_{\gamma(a,b)}(x) dx = \frac{a}{b^2}. \quad (7.10)$$

7.3. Неравенство Рао–Крамера–Фреше и измерение эффективности оценок

В п. 7.1.5 введено понятие эффективности оценки неизвестного параметра Θ , которое определяется средним квадратом отклонения оценки от истинного значения параметра, т.е. величиной $E(\hat{\Theta} - \Theta)^2$. В связи с этим возникает вопрос: нельзя ли описать ту границу эффективности, т.е. тот минимум (по всем возможным оценкам $\hat{\Theta}$) среднего квадрата $E(\hat{\Theta} - \Theta)^2$, улучшить которую невозможно? Этот минимум и явился бы тогда той точкой начала отсчета эффективности оценки, отправляясь от которой можно было бы ввести абсолютную шкалу измерения эффективности оценок. На этот вопрос дает ответ неравенство Рао–Крамера–Фреше, известное также как *неравенство информации*.

Рассмотрим класс всевозможных оценок $\hat{\theta}$ *скалярного параметра* θ , от которого зависит плотность вероятности $f(x; \theta)$ исследуемой генеральной совокупности. Пусть

$$E \hat{\theta} = \int \hat{\theta}(x_1, \dots, x_n) L(x_1, \dots, x_n; \theta) dx_1 \dots dx_n = \theta + b_{\hat{\theta}}(\theta), \quad (7.11)$$

т.е. величина $b_{\hat{\theta}}(\theta)$ дает смещение оценки $\hat{\theta}$ (очевидно, если оценка θ несмещенная, то $b_{\hat{\theta}}(\theta) \equiv 0$).

Если плотность $f(x; \theta)$ удовлетворяет некоторым *условиям регулярности* (в смысле характера ее зависимости от параметра θ), а именно:

а) область возможных значений исследуемой случайной величины, в которой $f(x; \theta) \neq 0$, *не зависит от θ* ;

б) в формуле (7.11) и в тождестве $\int L(x_1, \dots, x_n; \theta) dx_1 \dots dx_n \equiv 1$ допустимо дифференцирование по θ под знаком интеграла;

в) величина $I(\theta; x)$, определенная соотношением (7.6'), не равна нулю; тогда для любой оценки $\hat{\theta}$ неизвестного параметра θ имеет место сле-

дующее неравенство:

$$E(\hat{\theta} - \theta)^2 \geq \frac{(1 + db_{\hat{\theta}}(\theta)/d\theta)^2}{n E \left[(d \ln f(x; \theta)/d\theta)^2 \right]}, \quad (7.12)$$

или, что то же,

$$E(\hat{\theta} - \theta)^2 \geq \frac{(1 + db_{\hat{\theta}}(\theta)/d\theta)^2}{n I(\theta; x)} \quad (7.12')$$

(доказательство этого факта см. в Приложении к гл. 7 на стр. 281).

Имеется обобщение неравенства (7.12) на случай k -мерного параметра $\Theta = (\theta^{(1)}, \dots, \theta^{(k)})$, $k \geq 2$. В этом случае при тех же условиях регулярности а)-в) для любой векторной *несмещенной* оценки $\hat{\Theta}$ неизвестного параметра Θ матрица

$$\Sigma(\hat{\Theta}) - \frac{1}{n} I^{-1}(\Theta, x) \quad (7.13)$$

является неотрицательно-определенной. Здесь $\Sigma(\hat{\Theta})$ — ковариационная матрица векторной оценки $\hat{\Theta} = (\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(k)})$, а $I^{-1}(\Theta, x)$ — матрица, обратная к информационной матрице, определенной соотношениями (7.7) при $X = x$ (т.е. при *единственном* наблюдении x).

Неравенства информации (7.12)–(7.13) дают возможность ввести количественную меру эффективности оценок в классе регулярных (в смысле соблюдения условий а)-в)) генеральных совокупностей. Естественно, в частности, измерять степень эффективности скалярной несмещенной оценки $\hat{\theta}$ неизвестного значения параметра θ отношением $e(\hat{\theta})$ минимально возможной величины дисперсии оценки, определяемой правой частью неравенства (7.12), к дисперсии данной конкретной оценки $\hat{\theta}$, т.е.

$$e(\hat{\theta}) = \frac{1}{n I(\theta; x)} : E(\hat{\theta} - \theta)^2. \quad (7.14)$$

Подсчитаем эффективность некоторых оценок параметров a и σ^2 в условиях примера 7.1.

1. Рассмотрим в качестве оценки среднего значения a нормальной случайной величины среднюю арифметическую (выборочное среднее), т.е. положим

$$\hat{a} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Так как наблюдения x_i независимы и одинаково распределены, имеем:

$$E\hat{a} = \frac{1}{n} \sum_{i=1}^n E x_i = \frac{1}{n} \sum_{i=1}^n a = a;$$

$$D\hat{a} = \frac{1}{n^2} \sum_{i=1}^n D x_i = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}.$$

Поскольку (см. (7.8)) $I(a, x) = 1/\sigma^2$, то в соответствии с (7.14) получаем, что $e(\bar{x}) = 1$, т. е. оценка $\bar{x}$ в данном случае является *неулучшаемой*.

2. Рассмотрим в качестве оценки дисперсии σ^2 нормальной случайной величины «подправленную» на несмещенность выборочную дисперсию

$$\bar{s}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Выше получено, что $E\bar{s}^2 = \sigma^2$, т. е. $\bar{s}^2$ является несмещенной оценкой дисперсии σ^2 . В п. 7.1.5 было показано, что

$$D\bar{s}^2 = \frac{2}{n-1} \sigma^4.$$

Поскольку (см. (7.9)) $I(\sigma^2, x) = 1/2\sigma^4$, то в соответствии с (7.14) получаем $e(\bar{s}^2) = (n-1)/n$, т. е. оценка $\bar{s}^2$ не является эффективной, хотя и близка к ней при больших объемах выборок. В то же время можно показать, что если в качестве оценки σ^2 взять статистику

$$s_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - a)^2,$$

что в данных условиях допустимо, так как среднее значение a считается известным, то она окажется точно эффективной.

Подчеркнем в заключение, что *информационное неравенство справедливо лишь в классе регулярных (в смысле соблюдения условий а)–в) § 7.3) генеральных совокупностей*. В частности, если область возможных значений исследуемой случайной величины, для которых плотность $f(x; \theta)$ положительна, *зависит от оцениваемого параметра*, то неравенство информации «не работает». Именно такими нерегулярными плотностями являются, например, равномерное распределение (в котором параметрами служат концы диапазона изменения соответствующей случайной величины, см. п. 3.1.7) и экспоненциальное распределение со сдвигом θ , т. е.

распределение, задаваемое плотностью

$$f(x; \theta) = \begin{cases} e^{-(x-\theta)} & \text{при } x \geq \theta; \\ 0 & \text{при } x < \theta. \end{cases}$$

Если, не обращая внимания на то, что эта плотность не удовлетворяет условиям а)-в), вычислить по формуле (7.6) количество информации, содержащейся в n независимых наблюдениях, то получим $I(\theta; x_1, \dots, x_n) = n$. Следовательно, в соответствии с информационным неравенством (7.12) мы должны были бы прийти к выводу, что дисперсия любой оценки $\hat{\theta}$ параметра θ не может быть меньше $1/n$. В то же время нетрудно вычислить (см. ниже, (пример 7.6)), что для оценки

$$\hat{\theta} = x_{\min}(n) - \frac{1}{n}, \quad (7.15)$$

где, как обычно, $x_{\min}(n)$ — это минимальное значение в выборке $x_1, \dots, x_n$, мы имеем:

$$E\hat{\theta} = \theta; \quad D\hat{\theta} = \frac{1}{n^2}.$$

Так что если для измерения эффективности оценки (7.15) воспользоваться формулой (7.14), то получим, что эффективность оценки $\hat{\theta}$ не просто больше единицы, но и стремится к бесконечности по мере роста объема выборки n (так как $e(\hat{\theta}) = \frac{1}{n} : \frac{1}{n^2} = n$). В подобных ситуациях оценки называют иногда «сверхэффективными».

Замечание о дискретных случайных величинах. Все изложенные выше результаты (понятие количества информации, неравенство информации, измерение эффективности оценки) распространяются на случай дискретных признаков при соблюдении тех же ограничений а)-в) с помощью внесения очевидных видоизменений: плотности $f(x; \theta)$ заменяются вероятностями $p_i(\theta) = P\{\xi = x_i^0 \mid \theta\}$, а интегрирование — суммированием по всем возможным значениям анализируемой дискретной случайной величины. Таким образом, в качестве дискретных аналогов количества информации (7.6) и информационного неравенства (7.12) будем иметь:

$$I(\theta; x_1, \dots, x_n) = n I(\theta; x) = n \sum_i \left(\frac{d \ln p_i(\theta)}{d\theta} \right)^2 p_i(\theta); \quad (7.16)$$

$$E(\hat{\theta} - \theta)^2 \geq \frac{(1 + db(\theta)/d\theta)^2}{n \sum_i (d \ln p_i(\theta)/d\theta)^2 p_i(\theta)}. \quad (7.17)$$

В качестве примера рассмотрим случайную величину ξ , подчиненную распределению Пуассона, т. е.

$$p_i(\lambda) = P\{\xi = i \mid \lambda\} = \frac{\lambda^i}{i!} e^{-\lambda}, \quad i = 0, 1, 2, 3, \dots$$

$$I(\lambda; x) = \sum_{i=0}^{\infty} \left(\frac{d \ln p_i(\lambda)}{d\lambda} \right)^2 p_i(\lambda) = \sum_{i=0}^{\infty} \left(\frac{i}{\lambda} - 1 \right)^2 p_i = \frac{1}{\lambda},$$

т. е. дисперсия любой несмещенной оценки $\hat{\lambda}$ параметра λ не может быть меньше, чем $1/n I(\lambda, x) = \lambda/n$. Если рассмотреть в качестве оценки λ выборочное среднее $\bar{x}$, то будем иметь:

$$\begin{aligned} E \hat{\lambda} &= E \left(\frac{x_1 + \dots + x_n}{n} \right) = \frac{n E x}{n} = \lambda; \\ D \hat{\lambda} &= D \left(\frac{x_1 + \dots + x_n}{n} \right) = \frac{1}{n^2} n D x = \frac{\lambda}{n}. \end{aligned} \quad (7.18)$$

Таким образом, оценка $\hat{\lambda} = \bar{x}$ параметра λ в распределении Пуассона является эффективной.

7.4. Понятие об интервальном оценивании и доверительных областях (постановка задач)

Вычисляя на основании имеющихся у нас выборочных данных оценку $\hat{\theta}(x_1, \dots, x_n)$ параметра θ , мы отдаем себе отчет в том, что на самом деле величина $\hat{\theta}$ является лишь *приближенным* значением неизвестного параметра θ даже в том случае, когда эта оценка состоятельна (т. е. стремится к θ с ростом n), несмещенна (т. е. совпадает с θ в среднем) и эффективна (т. е. обладает наименьшей степенью случайных отклонений от θ). Возникает вопрос: как сильно может отклоняться это приближенное значение (оценка) от *истинного*? В частности, нельзя ли указать такую величину Δ , которая с «практической достоверностью» (т. е. с заранее заданной вероятностью, близкой к единице) гарантировала бы выполнение неравенства $|\hat{\theta} - \theta| < \Delta$? Или, что то же, нельзя ли указать интервал вида $(\hat{\theta}_1, \hat{\theta}_2)$, который с заранее заданной вероятностью (близкой к единице) покрывал бы неизвестное нам истинное значение θ искомого параметра? При этом заранее выбираемая исследователем вероятность, близкая к единице, обычно называется *доверительной вероятностью*, а сам интервал

$(\hat{\theta}_1, \hat{\theta}_2)$ — доверительным интервалом (или *интервальной оценкой*, в отличие от *точечных* оценок $\hat{\theta}$). Доверительный интервал по своей природе *случаен* (потому и идет речь о *вероятности накрыть* некоторую не известную нам, но не случайную точку θ) как по своему расположению, так и по своей длине. Величины $\hat{\theta}_1, \hat{\theta}_2$ и Δ , как правило, тоже строятся как функция выборочных данных $x_1, \dots, x_n$. Ширина доверительного интервала существенно зависит от объема выборки n (уменьшается с ростом n) и от величины доверительной вероятности (увеличивается с приближением доверительной вероятности к единице).

Все данные здесь определения и понятия без труда переносятся на случай *векторного* параметра $\Theta = (\theta^{(1)}, \dots, \theta^{(k)})$ с заменой доверительного интервала *доверительной областью* в соответствующем k -мерном пространстве.

7.5. Методы статистического оценивания неизвестных параметров

В предыдущем пункте рассмотрены различные варианты использования функций от исходных наблюдений $x_1, x_2, \dots, x_n$ в качестве оценок неизвестных параметров, анализировались их свойства. Однако пока не ясно, каким способом устанавливаются именно те комбинации результатов наблюдений (статистики), с помощью которых производится (да еще наилучшим в определенном смысле образом!) оценивание того или иного параметра. Каким образом, например, было установлено, что именно комбинации $\bar{x}, s^2$ лучше всего использовать в качестве оценок неизвестных параметров соответственно среднего значения $a = E\xi$ и дисперсии $\sigma^2 = D\xi$ нормальной генеральной совокупности? И как конкретно строить описанные выше доверительные интервалы и области для неизвестных значений параметров?

Описанию основных приемов, позволяющих получать ответы на данные вопросы, и посвящен настоящий пункт.

7.5.1. Метод максимального (наибольшего) правдоподобия

В соответствии с этим методом оценка $\hat{\Theta}_{\text{мп}}$ неизвестного параметра Θ по наблюдениям $x_1, \dots, x_n$ случайной величины ξ (подчиненной закону распределения $f_\xi(x, \Theta)$, где f — плотность или вероятность $P\{\xi = x\}$) определяется из условия

$$L(x_1, \dots, x_n; \hat{\Theta}_{\text{мп}}) = \max_{\hat{\Theta}} L(x_1, \dots, x_n; \hat{\Theta}), \quad (7.19)$$

где L — функция правдоподобия, определенная соотношением (7.5).

Таким образом, в формальной записи *оценка максимального правдоподобия* $\hat{\Theta}_{\text{мп}}$ параметра Θ по *независимым* наблюдениям $x_1, \dots, x_n$ может быть представлена в виде

$$\hat{\Theta}_{\text{мп}} = \arg \max_{\Theta} \prod_{i=1}^n f(x_i; \Theta). \quad (7.19')$$

Естественность подобного подхода к определению статистических оценок вытекает из смысла функции правдоподобия. Действительно, по определению (см. п. 7.2), функция $L(x_1, \dots, x_n; \Theta)$ при каждом фиксированном значении параметра Θ является мерой правдоподобности получения системы наблюдений, равных $x_1, x_2, \dots, x_n$. Поэтому, изменяя значения параметра Θ при данных *конкретных* (имеющихся у нас) величинах $x_1, x_2, \dots, x_n$, мы можем проследить, при каких значениях Θ эти наблюдения являются *более правдоподобными*, а при каких — менее и выбрать в конечном счете такое значение параметра $\hat{\Theta}_{\text{мп}}$, при котором имеющаяся у нас система наблюдений $x_1, \dots, x_n$ выглядит *наиболее правдоподобной* (очевидно, что это значение $\hat{\Theta}_{\text{мп}}$ определяется конкретными величинами наблюдений $x_1, x_2, \dots, x_n$, т.е. является некоторой функцией от них).

Так, например, пусть ξ — заработная плата работников, подчиненная логарифмически нормальному распределению (см. п. 3.1.6). И пусть с целью приближенной оценки средней величины логарифма заработной платы работников $a = E(\ln \xi)$ мы зафиксировали значения заработной платы работников $x_1 = 190$ ден.ед., $x_2 = 175$ ден.ед. и $x_3 = 205$ ден.ед. у трех случайно отобранных из интересующей нас совокупности работников. Тогда, расположив $y_i = \ln x_i$ ($i = 1, 2, 3$) на оси возможных значений нормально распределенной случайной величины $\eta = \ln \xi$, мы будем стараться подобрать такое значение $\hat{a}_{\text{мп}}$ параметра a в (a, σ^2) -нормальном распределении, при котором наши наблюдения y_1, y_2, y_3 выглядели бы *наиболее правдоподобными*, а именно, при котором произведение трех ординат плотности $\varphi(y; a; \sigma^2)$, вычисленных в точках соответственно $y_1 = \ln 190 = 5,25$, $y_2 = \ln 175 = 5,16$ и $y_3 = \ln 205 = 5,32$, достигало бы своего максимального значения.

На рис. 7.2 изображены графики функции плотности $\varphi(y; a; \sigma^2)$ при значении параметра $\hat{a}_{\text{мп}} = \bar{y} = 5,243$, соответствующем наибольшей правдоподобности наблюдений $y_1 = 5,25$, $y_2 = 5,16$ и $y_3 = 5,32$ (сплошная кривая), и при значении параметра $\hat{a} = 5,443$, при котором наши наблюдения выглядят явно неправдоподобными, — пунктирная кривая (значение дисперсии σ^2 определено в обоих случаях с помощью подправленной на несмещенность оценки максимального правдоподобия и равно 0,0064).

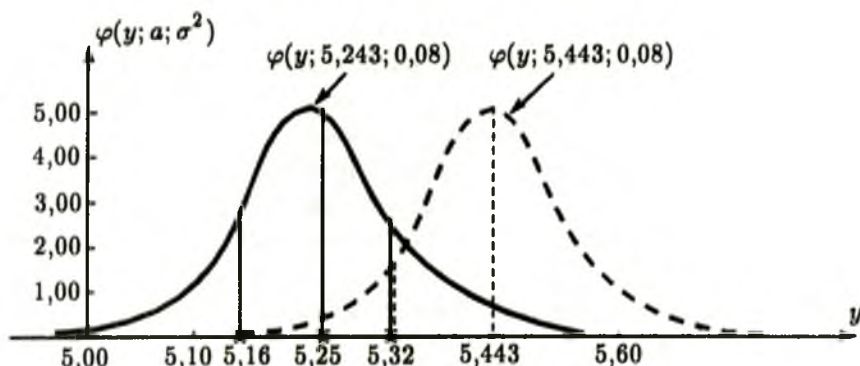


Рис. 7.2. Графики нормальной функции плотности при двух значениях параметра a

Отмеченная естественность подхода, исходящего из максимальной правдоподобности имеющихся наблюдений, подкрепляется хорошими свойствами оценок, получаемых с его помощью. Можно показать, в частности, что при достаточно широких условиях регулярности, накладываемых на изучаемый закон распределения $f(x; \Theta)$, оценки максимального правдоподобия $\hat{\Theta}_{\text{мп}}$ параметра Θ являются *состоятельными*, *асимптотически несмещенными* (т. е. их смещения стремятся к нулю при неограниченном увеличении объема выборки), *асимптотически нормальными* и *асимптотически эффективными* (т. е. их ковариационная матрица $\Sigma(\hat{\Theta}_{\text{мп}})$ асимптотически имеет вид $\Sigma(\hat{\Theta}_{\text{мп}}) = n^{-1} \times I^{-1}(\Theta; x)$, где $I(\Theta; X)$ — информационная матрица Фишера, определенная соотношениями (7.7) применительно к единственному наблюдению, т. е. при $X = x$).

Однако из этого не следует, что оценки максимального правдоподобия будут наилучшими во всех ситуациях. Во-первых, их хорошие свойства проявляются часто лишь при очень больших объемах выборок (т. е. являются асимптотическими), так что при малых n с ними могут конкурировать (и даже превосходить их) другие оценки, например, оценки метода моментов, метода наименьших квадратов и т. д. Во-вторых, и это, пожалуй, главное «узкое место» данного подхода, для построения оценок максимального правдоподобия и обеспечения их хороших свойств *необходимо точное знание типа анализируемого закона* распределения $f(x; \Theta)$, что в большинстве случаев оказывается практически нереальным. В подобных ситуациях бывает выгоднее искать не оценку, являющуюся наилучшей в рамках данного конкретного общего вида $f(x; \Theta)$ (но, как часто бывает, резко теряющую свои хорошие свойства при отклонениях реаль-

ного распределения от типа $f(x; \Theta)$), а оценку, хотя и не наилучшую в рамках совокупности $f(x; \Theta)$, но обладающую *достаточно устойчивыми* свойствами в более широком классе распределений, включающем в себя $f(x; \Theta)$ в качестве частного случая (см. ниже, п. 7.5.4). Подобные оценки принято называть *устойчивыми* или *робастными* (английский термин *robust estimation* означает грубое, или устойчивое оценивание). И наконец, оценки максимального правдоподобия *могут не быть даже состоятельными*, если число k оцениваемых по выборке параметров $\theta_1, \dots, \theta_k$ велико (имеет тот же порядок, что и объем выборки n) и растет вместе с увеличением числа наблюдений. Ниже приведен пример подобной ситуации (см. пример 7.6).

Попытаемся ответить на вопрос, как конкретно находятся оценки максимального правдоподобия, т.е. как проводится решение оптимизационной задачи типа (7.19').

Если функция $f(x; \Theta)$ удовлетворяет определенным условиям регулярности (дифференцируемость по Θ и т.п., см. условия а)–в) п. 7.3) и экстремум в (7.19) достигается во внутренней точке области допустимых значений неизвестного параметра Θ , то в точке $\hat{\theta}_{\text{мп}}$ должны обращаться в нуль частные производные функции $L(x_1, \dots, x_n; \Theta)$, а следовательно, и *логарифмической функции правдоподобия*

$$l(x_1, \dots, x_n; \Theta) = \ln L(x_1, \dots, x_n; \Theta) = \sum_{i=1}^n \ln f(x_i, \Theta) \quad (7.20)$$

в силу монотонного характера этой зависимости: последняя удобнее для вычислений. Значит, в данном случае оценка максимального правдоподобия $\hat{\Theta}_{\text{мп}} = (\hat{\theta}_{\text{мп}}^{(1)}, \dots, \hat{\theta}_{\text{мп}}^{(k)})$ должна удовлетворять уравнениям:

$$\frac{\partial l(x_1, \dots, x_n; \Theta)}{\partial \theta^{(j)}} = 0, \quad j = 1, 2, \dots, k, \quad (7.21)$$

и может определяться в качестве решения этой системы уравнений.

Однако могут быть ситуации (случай нерегулярных по Θ законов распределения), когда система (7.21) не определена или не имеет решений, в то время как решение (7.19') существует. В подобных ситуациях оценку $\hat{\Theta}_{\text{мп}}$ следует искать другими способами, в том числе с помощью непосредственного подбора решения (7.19).

Пример 7.2. Исследуемая случайная величина ξ имеет нормальную плотность вероятности

$$\varphi(x; a; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

с неизвестным средним значением $a = E\xi$ и неизвестной дисперсией $\sigma^2 = D\xi$.

В соответствии с (7.5) функция правдоподобия в этом случае будет

$$L(x_1, \dots, x_n; a; \sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - a)^2}.$$

Соответствующая логарифмическая функция правдоподобия

$$l(x_1, \dots, x_n; a, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - a)^2.$$

Дифференцируя l по a и σ^2 и последовательно приравнявая соответствующие частные производные к нулю, получаем конкретный вид системы (7.21):

$$\begin{aligned} \frac{\partial l(x_1, \dots, x_n; a, \sigma^2)}{\partial a} &= \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - a) = 0; \\ \frac{\partial l(x_1, \dots, x_n; a, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (x_i - a)^2 = 0. \end{aligned}$$

Решение этой системы относительно a и σ^2 дает оценки максимального правдоподобия этих параметров:

$$\hat{a}_{\text{мп}} = \frac{1}{n} \sum_{i=1}^n x_i; \quad \hat{\sigma}_{\text{мп}}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Выше (см. п. 7.3) установлено, что оценка $\hat{\theta}_{\text{мп}} = \bar{x}$ является эффективной оценкой параметра a (так как ее эффективность $e(\hat{a}_{\text{мп}}) = 1$), а оценка $\hat{\sigma}_{\text{мп}}^2 = s^2$ — асимптотически эффективной оценкой параметра σ^2 (так как ее эффективность $e(\hat{\sigma}_{\text{мп}}^2) = (n-1)/n$).

Пример 7.3. Исследуемая случайная величина ξ подчинена закону распределения Пуассона, т. е.

$$f(x; \lambda) = P\{\xi = x\} = \frac{\lambda^x}{x!} e^{-\lambda} \quad (x = 0, 1, 2, \dots),$$

с неизвестным значением параметра λ .

В соответствии с (7.20) логарифмическая функция правдоподобия, построенная по выборке $x_1, x_2, \dots, x_n$, имеет вид

$$\begin{aligned} l(x_1, \dots, x_n; \lambda) &= \sum_{i=1}^n (x_i \ln \lambda - \ln(x_i!) - \lambda) \\ &= (\ln \lambda) \cdot \sum_{i=1}^n x_i - \sum_{i=1}^n \ln(x_i!) - n\lambda. \end{aligned}$$

Отсюда после дифференцирования по λ получаем уравнение метода максимального правдоподобия

$$\frac{1}{\lambda} \sum_{i=1}^n x_i - n = 0,$$

откуда

$$\hat{\lambda}_{\text{мп}} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Легко видеть, что эта оценка несмещенная, так как

$$\mathbf{E} \hat{\lambda}_{\text{мп}} = \mathbf{E} \left(\sum_{i=1}^n x_i \right) / n = \left(\sum_{i=1}^n \mathbf{E} x_i \right) / n = n\lambda / n = \lambda.$$

Вычислим эффективность оценки $\hat{\lambda}_{\text{мп}}$. Нижняя граница дисперсии по всем возможным оценкам параметра λ может быть вычислена в соответствии с неравенством информации (7.12):

$$(\mathbf{D} \hat{\lambda})_{\min} = \frac{1}{n \mathbf{E} \left(\frac{d}{d\lambda} \ln f(x; \lambda) \right)^2} = \frac{1}{n \mathbf{E} \left(\frac{x}{\lambda} - 1 \right)^2} = \frac{\lambda^2}{n \mathbf{E} (x - \lambda)^2} = \frac{\lambda}{n}.$$

Дисперсию оценки $\hat{\lambda}_{\text{мп}}$ вычислим, опираясь на следующий известный факт: дисперсия суммы независимых случайных величин $x_1, x_2, \dots, x_n$ равна сумме их дисперсий, т. е.:

$$\mathbf{D} \hat{\lambda}_{\text{мп}} = \mathbf{D} \left(\frac{1}{n} \sum_{i=1}^n x_i \right) = \frac{1}{n^2} \sum_{i=1}^n \mathbf{D} x_i = \frac{1}{n^2} \cdot n\lambda = \frac{\lambda}{n}.$$

Сравнивая $(\mathbf{D} \hat{\lambda})_{\min}$ с $\mathbf{D} \hat{\lambda}_{\text{мп}}$, убеждаемся, что оценка максимального правдоподобия среднего значения пуассоновской случайной величины является эффективной.

Пример 7.4. Исследуемая случайная величина ξ распределена по равномерному закону (см. п. 3.1.7), т. е.

$$f(x; a, b) = \begin{cases} \frac{1}{b-a}, & \text{если } x \in [a, b]; \\ 0 & \text{— в противном случае,} \end{cases}$$

где параметры a и b неизвестны (подлежат оцениванию).

Легко проверить, что это — случай нерегулярный (в первую очередь потому, что область возможных значений исследуемого признака, в которой плотность положительна, зависит от оцениваемых по выборке параметров a и b). Поэтому обычная техника, использующая уравнения (7.21) метода максимального правдоподобия, здесь неприменима. Однако в этом случае экстремальная задача (7.19) может быть решена непосредственно.

Действительно,

$$L(x_1, \dots, x_n; a, b) = \frac{1}{(b-a)^n},$$

причем область допустимых значений параметров a и b , где производится поиск тех значений $\hat{a}_{\text{мп}}$ и $\hat{b}_{\text{мп}}$, при которых $1/(b-a)^n = \max_{a,b}$, описывается соотношениями:

$$a \leq \min_{1 \leq i \leq n} \{x_i\} = x_{\min}(n);$$

$$b \geq \max_{1 \leq i \leq n} \{x_i\} = x_{\max}(n),$$

где $x_1, x_2, \dots, x_n$ — имеющиеся в нашем распоряжении наблюдения исследуемой случайной величины. Очевидно, решение экстремальной задачи

$$\max \frac{1}{(b-a)^n}, \quad a \leq x_{\min}(n), \quad b \geq x_{\max}(n)$$

дается соотношениями:

$$\hat{a}_{\text{мп}} = x_{\min}(n); \quad \hat{b}_{\text{мп}} = x_{\max}(n).$$

Опираясь на результаты п. 6.3.1 (см. формулу (6.42)), можно подсчитать:

$$\begin{aligned} E\hat{a}_{\text{мп}} &= Ex_{\min}(n) = a + \frac{b-a}{n+1}; \\ D\hat{a}_{\text{мп}} &= Dx_{\min}(n) = \frac{n}{(n+1)^2(n+2)}(b-a)^2; \\ E\hat{b}_{\text{мп}} &= Ex_{\max}(n) = b - \frac{b-a}{n+1}; \\ D\hat{b}_{\text{мп}} &= Dx_{\max}(n) = \frac{n}{(n+1)^2(n+2)}(b-a)^2. \end{aligned} \tag{7.22}$$

Использовать неравенство информации для вычисления нижней границы дисперсии этих оценок мы не можем, так как случай нерегулярный. Из (7.22) видно, что величины $D\hat{a}_{\text{мп}}$ и $D\hat{b}_{\text{мп}}$ характеризуют одновременно средний квадрат отклонения «подправленных на несмещенность» оценок

$$\begin{aligned}\hat{a}'_{\text{мп}} &= x_{\min}(n) - \frac{x_{\max}(n) - x_{\min}(n)}{n-1}; \\ \hat{b}'_{\text{мп}} &= x_{\max}(n) + \frac{x_{\max}(n) - x_{\min}(n)}{n-1}\end{aligned}$$

от истинных значений параметров a и b .

Пример 7.5. Снова рассмотрим задачу оценивания параметра сдвига θ в экспоненциальном распределении, задаваемом плотностью

$$f(x; \theta) = \begin{cases} e^{-(x-\theta)} & \text{при } x \geq \theta; \\ 0 & \text{при } x < \theta. \end{cases}$$

Как и в предыдущем примере, имеем дело с *нерегулярным* случаем. Поэтому приходится непосредственно решать экстремальную задачу вида

$$\begin{aligned}\max_{\theta} L(x_1, \dots, x_n; \theta) &= \max_{\theta} e^{-\sum_{i=1}^n (x_i - \theta)}; \\ \theta &\leq \min_{1 \leq i \leq n} \{x_i\} = x_{\min}(n).\end{aligned}\tag{7.23}$$

Легко видеть, что $\hat{\theta}_{\text{мп}} = x_{\min}(n)$ является решением этой задачи: при любом другом θ , удовлетворяющем условию (7.23), очевидно

$$\begin{aligned}\sum_{i=1}^n (x_i - \theta) &= \sum_{i=1}^n [x_i - (\hat{\theta}_{\text{мп}} - \varepsilon)] = \sum_{i=1}^n (x_i - \hat{\theta}_{\text{мп}}) \\ &+ n\varepsilon > \sum_{i=1}^n (x_i - \hat{\theta}_{\text{мп}})\end{aligned}$$

и, следовательно,

$$L(x_1, \dots, x_n; \hat{\theta}_{\text{мп}}) > L(x_1, \dots, x_n; \theta).$$

Опираясь на результаты п. 6.3.1 (см. формулу (6.42)), можно подсчитать:

$$\begin{aligned}E\hat{\theta}_{\text{мп}} &= Ex_{\min}(n) = \theta + \frac{1}{n}; \\ D\hat{\theta}_{\text{мп}} &= E(x_{\min}(n) - Ex_{\min}(n))^2 = \frac{1}{n^2}; \\ E(\hat{\theta}_{\text{мп}} - \theta)^2 &= D\hat{\theta}_{\text{мп}} + \frac{1}{n^2} = \frac{2}{n^2}.\end{aligned}\tag{7.24}$$

Однако оценка $\hat{\theta}_{\text{мп}}^* = \hat{\theta}_{\text{мп}} - \frac{1}{n}$, получающаяся из оценки $\hat{\theta}_{\text{мп}}$ «подправлением на несмещенность», будет иметь средний квадрат отклонения

$$E(\hat{\theta}_{\text{мп}}^* - \theta)^2 = \frac{1}{n^2}. \quad (7.25)$$

Пример 7.6 (заимствован из [Ван дер Варден Б.Л., с. 187]). Рассмотрим ситуацию, когда метод максимального правдоподобия не приводит к состоятельной оценке.

С целью оценки n концентраций некоторого элемента $a_1, a_2, \dots, a_n$ в лаборатории производились двукратные измерения (x_i, y_i) каждой из концентраций a_i . Предполагается, что все $2n$ результатов измерений $x_1, y_1, x_2, y_2, \dots, x_n, y_n$ имеют одинаковую точность и являются независимыми нормальными случайными величинами (см. п. 3.1.5), так что в качестве функции правдоподобия получаем

$$\begin{aligned} L(x_1, y_1, \dots, x_n, y_n; a_1, a_2, \dots, a_n; \sigma^2) \\ = \frac{1}{(2\pi)^n \sigma^{2n}} \cdot e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n [(x_i - a_i)^2 + (y_i - a_i)^2]}. \end{aligned}$$

Неизвестными параметрами являются n средних значений $a_1, a_2, \dots, a_n$ и дисперсия σ^2 . Нетрудно получить оценки максимального правдоподобия параметров a_i :

$$\hat{a}_{\text{имп}} = \frac{1}{2}(x_i + y_i).$$

Решая теперь уравнение максимального правдоподобия (7.21), в которое вместо a_i подставлены значения $\hat{a}_{\text{имп}}$, получаем

$$\hat{\sigma}_{\text{мп}}^2 = \frac{1}{4n} \sum_{i=1}^n (x_i - y_i)^2.$$

Нетрудно подсчитать, что $E\hat{\sigma}_{\text{мп}}^2 = \sigma^2/2$, т.е. метод максимального правдоподобия дает в этом случае оценку параметра σ^2 с постоянным (асимптотически неустранимым) отрицательным смещением, равным $-\sigma^2/2$. В качестве наилучшей несмещенной оценки следовало бы выбрать в данном случае статистику

$$\hat{\sigma}^2 = \frac{1}{2n} \sum_{i=1}^n (x_i - y_i)^2.$$

7.5.2. Метод моментов

Пусть ξ — исследуемая одномерная случайная величина, подчиняющаяся закону распределения $f(x; \Theta)$, где функция $f(x; \Theta)$ — плотность вероятности, если ξ непрерывна, и вероятность $P\{\xi = x \mid \Theta\}$, если ξ дискретна, зависит от некоторого, вообще говоря, многомерного параметра $\Theta = (\theta^{(1)}, \dots, \theta^{(k)})$. И пусть мы хотим оценить неизвестное значение этого параметра, т. е. построить оценку $\hat{\Theta}$ по имеющейся в нашем распоряжении выборке, состоящей из независимых наблюдений $x_1, \dots, x_n$.

Метод моментов заключается в приравнивании определенного количества выборочных моментов к соответствующим теоретическим (т. е. вычисленным с использованием функции $f(x; \Theta)$) моментам исследуемой случайной величины, причем последние, очевидно, являются функциями от неизвестных параметров $\theta^{(1)}, \dots, \theta^{(k)}$. Рассматривая количество моментов, равное числу k подлежащих оценке параметров, и решая полученные уравнения относительно этих параметров, мы получаем искомые оценки. Таким образом, оценки $\hat{\theta}_{\text{мм}}^{(1)}, \dots, \hat{\theta}_{\text{мм}}^{(k)}$ по методу моментов неизвестных параметров $\theta^{(1)}, \dots, \theta^{(k)}$ являются решениями системы уравнений:

$$\int x^l \cdot f(x; \Theta) dx = \frac{1}{n} \sum_{i=1}^n x_i^l, \quad l = 1, 2, \dots, k; \quad (7.26)$$

(очевидно, если анализируемая случайная величина ξ дискретна, интегралы в левых частях (7.26) следует заменить соответствующими суммами).

Число уравнений в системах (7.26) должно быть равным числу k оцениваемых параметров. Вопрос о том, какие именно моменты включать в систему (7.26) (начальные, центральные или их некоторые модификации типа коэффициентов асимметрии или эксцесса), следует решать, руководствуясь конкретными целями исследования и сравнительной простотой формы зависимости альтернативных *теоретических* характеристик от оцениваемых параметров $\theta^{(1)}, \dots, \theta^{(k)}$. В статистической практике дело редко доходит даже до моментов четвертого порядка.

К достоинствам метода моментов следует отнести его сравнительно простую вычислительную реализацию, а также то, что оценки, полученные в качестве решений системы (7.26), являются функциями от выборочных моментов. Это упрощает исследование статистических свойств оценок метода моментов: можно показать (см. [Крамер Г., гл. 27 и 28]), что при довольно общих условиях распределение оценки такого рода при больших n асимптотически нормально, среднее значение такой оценки отличается от истинного значения параметра на величину порядка n^{-1} , а стандартное отклонение $\sigma(\hat{\theta}_{\text{мм}})$ асимптотически имеет вид $cn^{-1/2}$, где c —

некоторая постоянная величина.

В то же время, как показал Р. Фишер (Крамер Г.), асимптотическая эффективность оценок, полученных методом моментов, оказывается, как правило, меньше единицы, и в этом отношении они уступают оценкам, полученным методом максимального правдоподобия. Тем не менее метод моментов часто очень удобен на практике.

Иногда оценки, получаемые с помощью метода моментов, принимаются в качестве первого приближения, по которому можно определять другими методами оценки более высокой эффективности.

Вернемся к нашим примерам.

В примере 7.2 в качестве системы (7.26) имеем:

$$\begin{cases} a = \frac{1}{n} \sum_{i=1}^n x_i; \\ \sigma^2 + a^2 = \frac{1}{n} \sum_{i=1}^n x_i^2, \end{cases}$$

что дает уже знакомые нам по методу максимального правдоподобия оценки для параметров:

$$\begin{aligned} \hat{a}_{\text{мм}} &= \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}; \\ \hat{\sigma}_{\text{мм}}^2 &= \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = s^2. \end{aligned}$$

Нормальное распределение, так же как и распределение Пуассона (в чем легко убедиться, обратившись к примеру 7.3), относится к тем *редким* случаям, когда оценки по методу моментов совпадают с оценками по методу максимального правдоподобия.

Построение системы (7.26) в примере 7.4 дает:

$$\begin{cases} \frac{a+b}{2} = \bar{x}(n); \\ \frac{(b-a)^2}{12} = s^2(n). \end{cases}$$

Откуда легко получаем оценки:

$$\begin{aligned} \hat{a}_{\text{мм}} &= \bar{x}(n) - \sqrt{3}s(n); \\ \hat{b}_{\text{мм}} &= \bar{x}(n) + \sqrt{3}s(n). \end{aligned} \tag{7.27}$$

Можно сравнить асимптотическую эффективность оценок, полученных методом максимального правдоподобия и методом моментов: учитывая, что дисперсия оценок (7.27) как дисперсия функций выборочных моментов $\bar{x}(n)$ и $s^2(n)$ имеет порядок n^{-1} , и принимая во внимание соотношение (7.22), в соответствии с которым дисперсии оценок максимального правдоподобия тех же параметров имеют порядок n^{-2} , получаем, что эффективность $\hat{a}_{\text{мм}}$ и $\hat{b}_{\text{мм}}$ в сравнении с эффективностью $\hat{a}_{\text{мп}}$ и $\hat{b}_{\text{мп}}$ стремится к нулю при $n \rightarrow \infty$.

Реализация метода моментов в примере 7.5 дает

$$1 + \theta = \bar{x}(n).$$

Следовательно, $\hat{\theta}_{\text{мм}} = \bar{x}(n) - 1$.

Для подсчета среднего значения и дисперсии оценки $\hat{\theta}_{\text{мм}}$ воспользуемся следующими фактами:

а) случайную величину x_i , распределенную экспоненциально с параметром сдвига θ , можно интерпретировать как частный случай гамма-распределенной случайной величины с параметрами $a = 1$, $b = 1$ и с параметром сдвига θ (см. п. 3.2.5);

б) сумма n независимых случайных величин $x_1, x_2, \dots, x_n$, каждая из которых распределена по закону гамма с параметрами $a = 1$ и $b = 1$ и с параметром сдвига θ , подчиняется гамма-распределению с параметрами $a = n$, $b = 1$ и с тем же самым параметром сдвига θ (см. п. 3.2.5).

Поэтому

$$E\hat{\theta}_{\text{мм}} = E(\bar{x}(n) - 1) = \frac{n + n\theta}{n} - 1 = \theta;$$

$$E(\hat{\theta}_{\text{мм}} - \theta)^2 = D\hat{\theta}_{\text{мм}} = D(\bar{x}(n) - 1) = \frac{n}{n^2} = \frac{1}{n}.$$

Учитывая выражение (7.25) для среднего квадрата ошибки «подправленной» оценки по методу максимального правдоподобия $\hat{\theta}'_{\text{мп}}$ того же параметра θ , получаем

$$\frac{E(\hat{\theta}'_{\text{мп}} - \theta)^2}{E(\hat{\theta}_{\text{мм}} - \theta)^2} = \frac{\frac{1}{n^2}}{\frac{1}{n}} = \frac{1}{n} \rightarrow 0 \quad \text{при } n \rightarrow \infty,$$

т. е. и в этом случае асимптотическая эффективность оценки по методу моментов стремится к нулю.

7.5.3. Оценивание с помощью «взвешенных» статистик; цензурирование, урезание выборок и порядковые статистики как частный случай взвешивания

Выборочные моменты $\hat{m}_k$ всегда являются состоятельными оценками соответствующих теоретических моментов m_k , если последние существуют (см. п. 6.2). Однако не во всякой генеральной совокупности они являются наиболее эффективными оценками. Так, например, мы видели (см. п. 7.5.1 и 7.5.2), что эффективность оценки среднего значения исследуемой случайной величины с помощью выборочного среднего $\bar{x} = \hat{m}_1$ существенно зависит от типа анализируемой генеральной совокупности: для нормальной генеральной совокупности она равна единице (см. пример 7.2 в п. 7.5.1), а для совокупности, подчиненной равномерному закону распределения, существенно меньше единицы и сильно проигрывает в сравнении, например, с эффективностью оценки

$$\hat{a} = \frac{1}{2}(x_{\min} + x_{\max}). \quad (7.28)$$

Для построения оценки (7.28) нами использованы только два наблюдения из n имеющихся — наименьшее и наибольшее, т. е. оценка (7.28) относится к классу «взвешенных» порядковых статистик $\sum_{i=1}^n w_i x_{(i)}$, где $x_{(i)}$ — i -е по величине (в порядке возрастания) наблюдение, а w_i — его «вес» (очевидно, в статистике (7.28) принято $w_1 = w_n = 0,5$, а все остальные w_i равны нулю).

Выборочная медиана $\hat{x}_{\text{med}}$ демонстрирует удивительную устойчивость своих хороших свойств в качестве оценки теоретического среднего значения. Выборочная медиана $\hat{x}_{\text{med}}$ также относится к классу «взвешенных» порядковых статистик, т. е. статистик вида $\sum_{i=1}^n w_i x_{(i)}$; для ее получения в качестве частного случая статистик этого класса достаточно положить нулю все веса w_i , кроме одного ($w_{\frac{n+1}{2}} = 1$, если n нечетно) или кроме двух ($w_{\frac{n}{2}} = w_{\frac{n}{2}+1} = \frac{1}{2}$, если n четно).

Остановимся далее на описании основных подходов, связанных с использованием взвешенных статистик, и на классификации их типов.

Взвешивание выборочных данных $x_1, \dots, x_n$. В общем случае наблюдению x_i приписывается вес $w_i = w(x_i) \geq 0$, который определяется как некоторая функция от его текущего значения. Обычно веса подчиняют условию нормировки $\sum_{i=1}^n w(x_i) \equiv 1$. В частности, можно рассматривать w -взвешенные моменты случайной величины ξ с плотностью $f_\xi(x)$, как

выборочные $\hat{m}_k(n, W)$, так и теоретические $m_k(W)$:

$$\hat{m}_k(n, W) = \sum_{i=1}^n x_i^k \cdot w(x_i);$$

$$m_k(W) = \int v^k w(x) f_x(x) dx.$$

Под W понимается **вектор** весов $(w(x_1), \dots, w(x_n))$ в выражении для **выборочных** моментов и **функция** со значениями $w(x)$ в выражении для теоретических моментов.

Если имеют дело с результатами наблюдения одномерной случайной величины $x_1, \dots, x_n$, то часто вес наблюдения x_i определяют в зависимости от его порядкового номера в упорядоченном (по возрастанию) ряду наблюдений, т. е. располагают наблюдения в **вариационный ряд** $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ и каждому члену вариационного ряда $x_{(i)}$ ставят в соответствие некоторый вес w_i .

Примеры такого рода взвешивания (которое приводит к так называемым порядковым статистикам) приведены выше.

Цензурирование выборки. Этот прием заключается в приписывании ряду «хвостовых» членов вариационного ряда **нулевых весов**, а остальным — одинаковых положительных. Если приписывание нулевых весов производится по признаку выхода текущих значений наблюдений за пределы заданного диапазона $[a, b]$, т. е.

$$w(x_{(i)}) = \begin{cases} w_0 > 0, & \text{если } a \leq x_{(i)} \leq b; \\ 0, & \text{если } x_{(i)} < a \text{ или } x_{(i)} > b, \end{cases}$$

то говорят о **цензурировании 1-го типа**. Очевидно, в этом случае число ν оставшихся в рассмотрении наблюдений есть величина **случайная** ($\nu < n$).

Если же нулевые веса приписываются **фиксированной доле α** крайних малых значений и **фиксированной доле β** крайних больших значений, то говорят, что производится **цензурирование 2-го типа уровня (α, β)** . В этом случае число ν оставшихся в рассмотрении наблюдений является величиной, заранее заданной и равной, в частности, $n(1 - \alpha - \beta)$.

Исследователь может прибегнуть к цензурированию **вынужденно** или **добровольно**. Вынужденное цензурирование обусловлено соответствующими условиями эксперимента: например, мы ставим на разрушающие испытания n изделий, но можем производить эксперимент в течение ограниченного времени T . Очевидно, мы будем вынуждены произвести в данном случае одностороннее цензурирование 1-го типа, при котором из дальнейшего рассмотрения исключаются точные значения долговечностей

(времени до разрушения) всех тех изделий, которые не разрушились за время T . С другой стороны, в классе оценок, построенных по цензурированным выборкам, часто можно найти оценки, хотя и не являющиеся наилучшими в жестких рамках генеральной совокупности определенного типа, но обладающие выгодными свойствами устойчивости своих хороших качеств по отношению к тем или иным отклонениям от априорных допущений.

Урезание распределения. Это понятие связано с ситуациями, когда исследуемый признак ξ просто не может быть наблюдаем в какой-либо части области его возможных значений. Так, например, если мы исследуем распределение семей по доходу, но по условиям выборочного обследования лишены возможности наблюдать семьи со среднедушевым доходом, меньшим некоторого заданного уровня a (тыс. руб.), то в подобных случаях говорят, что *распределение урезано слева в точке a* . В отличие от цензурированных выборок в выборках из урезанных распределений мы не имеем возможности оценить даже доли наблюдений, располагающихся за пределами порога урезания.

Весьма подробные сведения об использовании в задачах статистического оценивания параметров взвешенных и, в частности, порядковых статистик и статистик, построенных по цензурированным выборкам, с обсуждением различных вопросов устойчивости получаемых при этом оценок читатель найдет, например, в [Кендалл М. Дж., Стьюарт А., 1973, гл. 32] и [Дейвид Г.].

7.5.4. Построение интервальных оценок (доверительных областей)

В п. 7.4 введено понятие интервальной оценки неизвестного параметра $\Theta = (\theta^{(1)}, \dots, \theta^{(k)})^T$, которую называют также доверительным интервалом, а при многомерном параметре, т.е. при $k \geq 2$, — доверительной областью. Как же конкретно построить по выборочным данным $x_1, \dots, x_n$ такую случайную область $\Delta\Theta_P(x_1, \dots, x_n)$, которая с наперед заданной доверительной вероятностью P накрывала бы неизвестное нам значение параметра Θ ? Очевидно, эта область должна конструироваться вокруг *точечной* оценки $\hat{\Theta}$ параметра Θ , а ее точный вид и объем определяются характером закона распределения случайной величины $\hat{\Theta}$, в частности, ее функцией распределения, которая, к сожалению, тоже зависит от неизвестного истинного значения параметра Θ .

Существуют два подхода к преодолению этой трудности. **Первый** подход, если его удастся реализовать, приводит к построению *точ-*

ны (при каждом конечном объеме выборки n) доверительных областей $\Delta\Theta_P(x_1, \dots, x_n)$. А реализовать его удастся в тех случаях, когда существует принципиальная возможность подбора такой функции от результатов наблюдения $x_1, x_2, \dots, x_n$ (т.е. такой статистики), закон распределения вероятностей которой обладал бы одновременно следующими свойствами:

- (а) не зависит от оцениваемого параметра Θ ;
- (б) описывается одним из стандартных затабулированных распределений (стандартным нормальным, χ^2 -, F -, Стьюдента);
- (в) из того факта, что значения данной статистики заключены в определенных пределах с заданной вероятностью, можно сделать вывод, что оцениваемый параметр тоже должен лежать между некоторыми границами с той же самой вероятностью.

При этом в выражении самой этой статистики анализируемый параметр Θ и его состоятельная точечная оценка $\hat{\Theta}$ (полученная одним из известных методов — максимального правдоподобия, моментов и т.д.) обычно участвуют в комбинациях разности ($\hat{\Theta} - \Theta$) или отношения ($\hat{\Theta}/\Theta$).

В качестве примера рассмотрим задачу интервального оценивания параметров a и σ^2 нормальной генеральной совокупности (см. пример 7.2 в п. 7.5.1).

Как известно (см. п. 6.30), статистика

$$\frac{\bar{x} - a}{s} \sqrt{n-1}$$

подчинена закону распределения Стьюдента с $n-1$ степенями свободы. Поэтому, определив из таблиц по заданной вероятности P процентные точки уровня $q = (1-P)/2$ и $1-q = (1-P)/2$ t -распределения с $n-1$ степенями свободы, т.е. $100q\%$ -ную точку $t_q(n-1)$ и $100(1-q)\%$ -ную точку $t_{1-q}(n-1)$ (причем в силу симметрии t -распределения $t_{1-q}(n-1) = -t_q(n-1)$), мы можем утверждать, что неравенство

$$-t_q(n-1) < \frac{(\bar{x} - a)\sqrt{n-1}}{s} < t_q(n-1)$$

выполняется с вероятностью $P = 1 - 2q$. А это означает, что случайный доверительный интервал

$$\Delta a_P(x_1, \dots, x_n) = \left[\bar{x} - t_q(n-1) \cdot \frac{s}{\sqrt{n-1}}; \bar{x} + t_q(n-1) \cdot \frac{s}{\sqrt{n-1}} \right] \quad (7.29)$$

накрывает неизвестное среднее значение a с заданной вероятностью P .

Для построения интервальной оценки параметра σ^2 воспользуемся тем фактом, что статистика $\frac{ns^2}{\sigma^2}$ подчинена χ^2 -распределению (хи-квадрат) с $n - 1$ степенями свободы (см. п. 6.27). Поэтому, определив из таблиц процентные точки χ^2 -распределения с $n - 1$ степенями свободы $\chi_{1-q}^2(n - 1)$ и $\chi_q^2(n - 1)$, где, как и прежде, $q = (1 - P)/2$, а P — заданная доверительная вероятность, имеем неравенство

$$\chi_{1-q}^2(n - 1) < \frac{ns^2}{\sigma^2} < \chi_q^2(n - 1),$$

которое выполняется с вероятностью $P = 1 - 2q$. Разрешая это неравенство относительно σ^2 , получаем, что случайный доверительный интервал

$$\Delta\sigma_P^2(x_1, \dots, x_n) = \left[\frac{ns^2}{\chi_q^2(n - 1)}; \frac{ns^2}{\chi_{1-q}^2(n - 1)} \right] \quad (7.30)$$

накрывает неизвестное значение дисперсии σ^2 с заданной вероятностью P .

Пример 7.7. Из многочисленного коллектива работников фирмы случайным образом отобрано $n = 25$ работников. Средняя заработная плата этих работников составила $\bar{x} = 700$ ден. ед. при среднеквадратическом отклонении $s = 100$ ден. ед. Предполагается, что распределение работников фирмы по размерам заработной платы подчиняется $(a; \sigma^2)$ -нормальному распределению. Требуется с доверительной вероятностью $P = 0,95$ определить интервальную оценку для:

- средней месячной заработной платы на фирме;
- суммы затрат фирмы на заработную плату отдела, состоящего из 520 сотрудников.

Решение.

а) Среднемесячная заработная плата на фирме характеризуется генеральной средней a . От нас требуется определить интервальную оценку a с доверительной вероятностью $P = 0,95$.

Согласно (7.29) имеем

$$a \in \left[\bar{x} \pm t_{0,025}(n - 1) \frac{s}{\sqrt{n - 1}} \right],$$

где $t_{0,025}(24) - 2,5\%$ -ная точка t -распределения.

По таблице процентных точек распределения Стьюдента приложения 1 находим для $q = \frac{1-P}{2} = 0,025$ и числа степеней свободы $\nu = n - 1 = 24$ процентную точку $t_{0,025}(24) = 2,064$. Тогда

$$a \in \left[700 \pm 2,064 \frac{100}{\sqrt{24}} \right].$$

Таким образом, с вероятностью $P = 0,95$ можно гарантировать, что средняя заработная плата на фирме находится в пределах: $657,88$ ден. ед. $\leq a \leq 742,12$ ден. ед.

б) Сумма затрат фирмы на заработную плату отдела составит Na ден. ед. Поэтому с вероятностью $P = 0,95$ можно гарантировать, что затраты фирмы на заработную плату не выйдут из интервала:

$$520 \cdot 657,88 \leq Na \leq 742,12 \cdot 520,$$

т. е.

$$342098 \text{ ден. ед.} \leq Na \leq 385902 \text{ ден. ед.}$$

Пример 7.8. При анализе точности фасовочного автомата было проведено $n = 24$ контрольных взвешиваний пятисотграммовых пачек кофе. Известно, что ошибки в расфасовке автомата подчиняются $(0; \sigma^2)$ -нормальному распределению (величина σ^2 — неизвестна). По результатам измерений рассчитано выборочное среднеквадратическое отклонение $s = 0,8$ г. Требуется с доверительной вероятностью $P = 0,95$ оценить точность фасовочного автомата, т. е. определить интервальную оценку σ .

Решение: Согласно (7.30) имеем интервальную оценку дисперсии

$$\sigma^2 \in \left[\frac{ns^2}{\chi_q^2}; \frac{ns^2}{\chi_{1-q}^2} \right],$$

где $q = \frac{1-P}{2} = 0,025$. По таблице процентных точек χ^2 -распределения (см. приложение 1) найдем:

$$\begin{aligned} \chi_q^2(n-1) &= \chi_{0,025}^2(23) = 38,0757; \\ \chi_{(1-q)}^2(n-1) &= \chi_{0,975}^2(23) = 11,6885. \end{aligned}$$

Тогда с вероятностью $0,95$:

$$\sigma^2 \in \left[\frac{24 \cdot 0,64}{38,0757}; \frac{24 \cdot 0,64}{11,6885} \right].$$

Отсюда с доверительной вероятностью $P = 0,95$ можно утверждать, что среднеквадратическое отклонение σ будет находиться в интервале

$$0,632 \text{ г.} \leq \sigma \leq 1,146 \text{ г.}$$

Предположив, что ошибка фасовочного автомата есть нормальная случайная величина с нулевой средней и среднеквадратическим отклонением σ ,

можно с вероятностью 0,954 утверждать, что вес пачек кофе будет в пределах $[500 - 2\sigma; 500 + 2\sigma] = [500 - 2,292; 500 + 2,292] \approx [497,71 \text{ г.}, 502,29 \text{ г.}]$.

Второй подход к построению доверительных областей более прост и универсален, однако он основан на *асимптотических* свойствах оценок, а поэтому дает приближенные результаты и пригоден лишь при достаточно больших объемах выборок n . Этот подход использует тот факт, что как оценки максимального правдоподобия, так и оценки по методу моментов имеют асимптотически нормальное совместное распределение, т. е. распределение k -мерного вектора $\xi(n) = \sqrt{n}(\hat{\Theta} - \Theta)$ стремится к многомерному нормальному закону с нулевым вектором средних значений и с ковариационной матрицей $\Sigma(\Theta)$, зависящей от неизвестного параметра Θ . При этом приближенном подходе допускаются две «натяжки»: во-первых, *асимптотический* вид распределений случайной величины $\xi(n)$ используется при *конечных* объемах выборки n и, во-вторых, вместо неизвестного значения параметра Θ в матрицу $\Sigma(\Theta)$ вставляется его оценочное значение $\hat{\Theta}$.

Теперь, для того чтобы построить доверительную область для неизвестного параметра $\Theta = (\theta^{(1)}, \dots, \theta^{(k)})^T$, мы должны воспользоваться следующим известным фактом (см. [Андерсон, с. 77]): если k -мерный вектор $\xi(n) = \sqrt{n}(\hat{\Theta} - \Theta)$ распределен нормально с параметрами 0 и $\Sigma(\Theta)$, то случайная величина

$$n(\hat{\Theta} - \Theta)^T \cdot \Sigma^{-1}(\Theta)(\hat{\Theta} - \Theta)$$

имеет χ^2 -распределение с k степенями свободы.

Определив из таблиц по заданной величине доверительной вероятности P процентные точки χ^2 -распределения с k степенями свободы $\chi^2_{1-q}(k)$ и $\chi^2_q(k)$, где $q = (1 - P)/2$, и заменив в известной матрице $\Sigma(\Theta)$ неизвестное значение параметра Θ его приближенным значением $\hat{\Theta}$, мы можем утверждать, что неравенство

$$\chi^2_{1-q}(k) < n(\hat{\Theta} - \Theta)^T \cdot \Sigma^{-1}(\hat{\Theta}) \cdot (\hat{\Theta} - \Theta) < \chi^2_q(k) \quad (7.31)$$

выполняется с вероятностью, приблизительно равной P .

З а м е ч а н и е 1. В случае *единственного* оцениваемого параметра θ (т. е. при $k = 1$) можно воспользоваться непосредственно $(0, \sigma_{\hat{\theta}}^2(\theta))$ -нормальностью разности $\hat{\theta} - \theta$ и записать вместо (7.31)

$$-u_q \cdot \sigma_{\hat{\theta}}(\theta) < \hat{\theta} - \theta < u_q \sigma_{\hat{\theta}}(\theta), \quad (7.31')$$

где u_q — 100 q %-ная точка стандартного нормального распределения, а $\sigma_{\hat{\theta}}^2(\theta)$ — дисперсия оценки $\hat{\theta}$. Из (7.31') следует запись соответствующего

доверительного интервала:

$$\Delta\theta_p(x_1, \dots, x_n) = [\hat{\theta} - u_q \cdot \sigma_{\hat{\theta}}(\theta), \hat{\theta} + u_q \cdot \sigma_{\hat{\theta}}(\theta)]. \quad (7.31'')$$

З а м е ч а н и е 2. Если в качестве Θ используются точечные оценки максимального правдоподобия, то ковариационная матрица $\Sigma(\Theta)$ вектора $\sqrt{n}(\hat{\Theta} - \Theta)$ однозначно определяется информационной матрицей Фишера (см. п. 7.2) $\Sigma(\theta) = I^{-1}(\theta)$, где элементы матрицы $I(\theta)$ определяются соотношениями (7.7).

З а м е ч а н и е 3. Положительная определенность и симметричность матрицы $\Sigma(\Theta)$ обуславливают *эллипсоидальный характер* доверительного множества, задаваемого соотношением (7.31).

П р и м е р 7.9. Рассмотрим задачу интервальной оценки по наблюдениям $x_1, \dots, x_n$ параметра p биномиального закона (см. п. 3.1.1), т. е. закона распределения дискретной случайной величины ξ , определяемого вероятностями

$$f(x; p, N) = P\{\xi = x\} = C_N^x p^x (1-p)^{N-x}, \quad x = 0, 1, \dots, N,$$

где N — известное целое положительное число, а p — параметр, подлежащий оценке ($0 < p < 1$).

Сначала в соответствии с техникой, описанной в п. 7.5.1, подсчитаем *точечную* оценку $\hat{p}$ максимального правдоподобия параметра p .

Логарифмическая функция правдоподобия в данном случае

$$\begin{aligned} l(x_1, x_2, \dots, x_n; p, N) &= \sum_{i=1}^n \ln C_N^{x_i} + \ln p \sum_{i=1}^n x_i \\ &\quad + \ln(1-p) \sum_{i=1}^n (N - x_i). \end{aligned}$$

Соответствующее уравнение максимального правдоподобия

$$\frac{dl}{dp} = \frac{1}{p} \sum_{i=1}^n x_i - \frac{1}{1-p} \sum_{i=1}^n (N - x_i) = 0.$$

Решая его относительно p , получаем оценку максимального правдоподобия:

$$\hat{p} = \frac{\sum_{i=1}^n x_i}{nN} = \frac{\bar{x}(n)}{N}.$$

Пользуясь независимостью x_i и тем фактом, что $E x_i = Np$ и $D x_i = Np(1-p)$ (см. п. 3.1.1), имеем:

$$E\hat{p} = p; \quad D\hat{p} = \frac{p(1-p)}{nN}.$$

Задавшись доверительной вероятностью $P = 0,95$, используя факт асимптотической нормальности разности $\hat{p} - p$ и подставляя в выражение для дисперсии $D\hat{p}$ вместо p его приближенное значение $\hat{p}$, получим в соответствии с (7.31'') интервальную оценку для p (с уровнем доверия 0,95):

$$\Delta p_{0,95}(x_1, \dots, x_n) = \left[\hat{p} - 1,96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{nN}}, \hat{p} + 1,96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{nN}} \right].$$

7.6. Байесовский подход к статистическому оцениванию

7.6.1. «Философия» байесовского подхода

Байесовский подход является одним из возможных способов формализации и операционализации тезиса, в справедливости которого нет видимых причин сомневаться: *степень нашей разумной уверенности в некотором утверждении (касающемся, например, оценки неизвестного численного значения интересующего нас параметра) возрастает и корректируется по мере пополнения имеющейся у нас информации относительно исследуемого явления*. Могут быть различные формы подтверждения этого тезиса, в том числе не имеющие отношения к байесовскому подходу. Одна из них выражена, например, в свойстве состоятельности оценки $\hat{\Theta}_n$ неизвестного параметра Θ : чем больше объем выборки n , на основании которой мы строим свою оценку $\hat{\Theta}_n$, тем большей информацией об этом параметре мы располагаем и тем ближе (в смысле сходимости $\hat{\Theta}_n$ к Θ по вероятности, см. выше, п. 7.1.3) к истине наше заключение.

Специфика именно байесовского способа операционализации этого тезиса основана на двух положениях.

1) Во-первых, «степень нашей разумной уверенности» в справедливости некоторого утверждения численно выражается в виде вероятности. Это означает, что вероятность в байесовском подходе выходит за рамки ее интерпретации в терминах условий статистического ансамбля (см. п. В.2.1 введения), но относится к одной из категорий субъективной школы теории вероятностей.

2) Во-вторых, статистик при принятии решения использует в качестве исходной информации одновременно информацию двух типов: *априорную и содержащуюся в исходных статистических данных* (см. п. В.3 введения). При этом априорная информация предоставлена ему в виде

некоторого *априорного распределения вероятностей* анализируемого неизвестного параметра, которое описывает степень его уверенности в том, что этот параметр примет то или иное значение, *еще до начала сбора исходных статистических данных*. По мере же поступления исходных статистических данных статистик уточняет (пересчитывает) это распределение, *переходя от априорного распределения к апостериорному*, используя для этого известную формулу Байеса (см. п. 1.1.3; отсюда и название подхода).

7.6.2. Общая логическая схема и базовые формулы байесовского метода оценивания параметров

Общая логическая схема байесовского метода оценивания неизвестных значений параметров представлена на рис. 7.3.

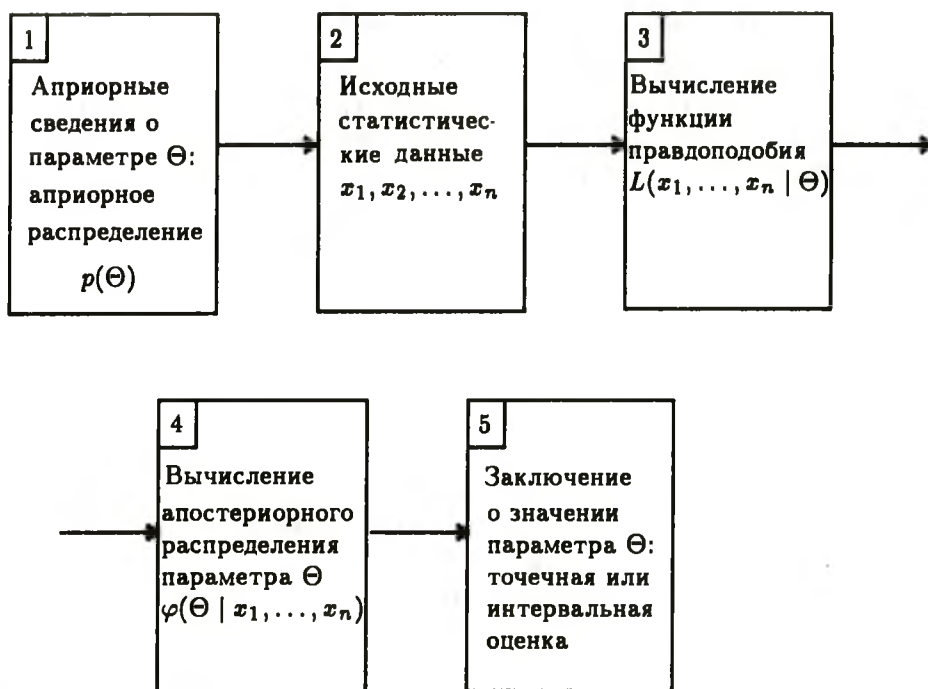


Рис. 7.3. Общая логическая схема байесовского подхода в статистическом оценивании

Рассмотрим реализацию схемы байесовского оценивания неизвестного параметра.

Априорные сведения о параметре Θ основаны на предыстории функционирования анализируемого процесса (если таковая имеется) и на профессиональных теоретических соображениях о его сущности, специфике, особенностях и т.п. В конечном итоге эти априорные сведения должны быть представлены в виде функции $p(\Theta)$, задающей *априорное распределение параметра* и интерпретируемой как вероятность того, что параметр примет значение, равное Θ , если параметр дискретен, и как функция плотности распределения в точке Θ , если параметр непрерывен по своей природе.

Примеры вероятностных законов, используемых в качестве априорных распределений параметров, будут приведены ниже (см. п. 7.6.3). Заметим лишь, что в ситуациях, когда априорные сведения об анализируемом параметре слишком скудны, в качестве априорного распределения $p(\Theta)$ используют, например, равномерное на отрезке $[\Theta_{\min}, \Theta_{\max}]$ распределение, где $[\Theta_{\min}, \Theta_{\max}]$ — априорный диапазон варьирования возможных значений оцениваемого параметра, т.е.:

$$p(\Theta) = \begin{cases} \frac{1}{\Theta_{\max} - \Theta_{\min}} & \text{при } \Theta_{\min} \leq \Theta \leq \Theta_{\max}; \\ 0 & \text{при } \Theta \notin [\Theta_{\min}; \Theta_{\max}]. \end{cases} \quad (7.32)$$

Исходные статистические данные $x_1, x_2, \dots, x_n$ — это не что иное, как выборка объема n из анализируемой генеральной совокупности. Получая исходные статистические данные, мы к имевшейся ранее априорной информации о параметре присоединяем выборочную (эмпирическую) информацию.

Вычисление функции правдоподобия $L(x_1, x_2, \dots, x_n | \Theta)$ производится в соответствии с (7.5) по формуле

$$L(x_1, x_2, \dots, x_n | \Theta) = f(x_1 | \Theta) \cdot f(x_2 | \Theta) \cdot \dots \cdot f(x_n | \Theta), \quad (7.33)$$

где $f(x | \Theta)$ — функция плотности (или вероятность $P\{\xi = x | \Theta\}$), описывающая закон распределения вероятностей анализируемой генеральной совокупности в предположении (или при условии), что значение оцениваемого параметра равно Θ .

Вычисление апостериорного распределения $\varphi(\Theta | x_1, \dots, x_n)$ осуществляется с помощью формулы Байеса (см. п. 1.1.3, формула (1.19)), в которой роль события A_i играет событие, заключающиеся в том, что значение оцениваемого параметра равно Θ , а роль условия B — события, заключающиеся в том, что значения n наблюдений, произведенных в анализируемой генеральной совокупности, зафиксированы на уровнях

$x_1, x_2, \dots, x_n$. Соответственно имеем:

$$\varphi(\Theta | x_1, \dots, x_n) = \frac{p(\Theta)L(x_1, \dots, x_n | \Theta)}{\int L(x_1, \dots, x_n | \Theta) \cdot p(\Theta) d\Theta}. \quad (7.34)$$

Построение байесовских точечных и интервальных оценок основано на использовании знания апостериорного распределения $\varphi(\Theta | x_1, \dots, x_n)$, задаваемого соотношением (7.34). В частности, в качестве байесовских точечных оценок $\hat{\Theta}^{(6)}$ используют среднее или модальное значение этого распределения, т. е.:

$$\hat{\Theta}_{\text{ср}}^{(6)} = E(\Theta | x_1, \dots, x_n) = \int \Theta \varphi(\Theta | x_1, \dots, x_n) d\Theta, \quad (7.35)$$

$$\hat{\Theta}_{\text{мод}}^{(6)} = \arg \max_{\Theta} \varphi(\Theta | x_1, \dots, x_n).$$

Отметим, что для вычисления этих оценок нам достаточно знать только числитель правой части (7.34), так как знаменатель этого выражения играет роль нормирующего множителя и от Θ не зависит (это существенно упрощает процесс практического построения оценок $\hat{\Theta}_{\text{ср}}^{(6)}$ и $\hat{\Theta}_{\text{мод}}^{(6)}$).

Отметим также одно важное оптимальное свойство оценки $\hat{\Theta}_{\text{ср}}^{(6)}$. Пусть $\hat{\Theta}(x_1, \dots, x_n)$ — любая оценка параметра Θ . Оказывается, если качество любой оценки $\hat{\Theta}(x_1, \dots, x_n)$ измерять так называемым *апостериорным байесовским риском*

$$\begin{aligned} R^{(6)}(x_1, \dots, x_n) &= E\{(\hat{\Theta}(x_1, \dots, x_n) - \Theta)^2 | x_1, \dots, x_n\} \\ &= \int (\hat{\Theta}(x_1, \dots, x_n) - \Theta)^2 \varphi(\Theta | x_1, \dots, x_n) d\Theta \end{aligned}$$

или его *средним* (усреднение — по всем возможным выборкам $x_1, \dots, x_n$) значением $R_{\text{ср}}^{(6)}$, то байесовская оценка (7.35) является наилучшей и в том и в другом смысле.

Для построения *байесовского доверительного интервала* для параметра Θ необходимо вычислить по формуле (7.34) функцию $\varphi(\Theta | x_1, \dots, x_n)$ апостериорного закона распределения параметра Θ , а затем по заданной доверительной вероятности P определить $100\frac{1+P}{2}$ - и $100\frac{1-P}{2}$ %-ные точки этого закона, которые и дают соответственно левый и правый концы искомой интервальной оценки.

Заметим, что байесовский способ оценивания может давать весьма ощутимый выигрыш в точности при ограниченных объемах выборок. В

процессе же неограниченного роста объема выборки n оба подхода будут давать, в силу их состоятельности, все более похожие результаты.

7.6.3. Примеры байесовского оценивания

Рассмотрим теперь на примерах, как практически реализуется описанная в предыдущем пункте схема байесовского оценивания неизвестного параметра.

Пример 7.10. Анализируется закон распределения вероятностей (з.р.в.) семей определенной социально-экономической страты по величине среднедушевого дохода ξ . Проведенные ранее исследования говорят о том, что з.р.в. среднедушевого дохода внутри однородной страты описывается *логарифмически нормальной* моделью (см. п. 3.1.6), а это значит, что случайная величина $\eta = \ln \xi$ подчинена нормальному закону. Требуется по результатам обследования среднедушевого дохода десяти семей, случайно отобранных из всей совокупности семей анализируемой страты, а также по некоторой априорной информации о величине теоретического среднего $\theta = E(\ln \xi)$ построить байесовскую оценку для θ .

Математическая постановка задачи. Мы располагаем следующей информацией об анализируемой генеральной совокупности.

1) Логарифм (натуральный) от величины среднедушевого дохода ($\eta = \ln \xi$) распределен нормально с *неизвестным* средним значением θ и *известной* дисперсией σ^2 , т. е.

$$f_{\eta}(x; \theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}; \quad \sigma^2 = 0,28.$$

2) Из предыстории и опыта обследований семей той же самой страты в других регионах известно, что величина θ ведет себя как $(\theta_0; \sigma_0^2)$ -нормальная случайная величина, где значения θ_0 и σ_0^2 известны, т. е.

$$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}}; \quad \theta_0 = 0,60; \quad \sigma_0^2 = 0,03.$$

3) Имеются результаты обследования n ($n = 10$) случайно отобранных от анализируемой страты семей по среднедушевому доходу, т. е. случайная выборка значений $x_1, x_2, \dots, x_n$, где $x_i = \ln y_i$, а y_i — среднедушевой доход i -й обследованной семьи. Ниже приводятся конкретные значения x_i :

$$x_1 = 0,54; \quad x_2 = 1,20; \quad x_3 = 0,36; \quad x_4 = 0,80; \quad x_5 = 0,42; \quad x_6 = 2,10; \\ x_7 = 0,70; \quad x_8 = 0,25; \quad x_9 = 0,90; \quad x_{10} = 0,48.$$

Требуется на основании исходной информации 1) – 3) построить байесовскую оценку для параметра θ и сравнить эту оценку с оценкой, полученной классическим методом максимального правдоподобия.

Решение. Поскольку мы уже располагаем априорным распределением $p(\theta)$ и исходными статистическими данными $x_1, \dots, x_n$, первые два этапа схемы, представленной на рис. 7.3, можно считать реализованными.

На 3-м этапе вычисляем функцию правдоподобия. В нашем случае имеем:

$$\begin{aligned} L(x_1, \dots, x_n | \theta) &= \prod_{i=1}^n f_{\eta}(x_i; \theta) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2 \right\} \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp \left\{ -\frac{1}{2\sigma^2} [ns^2 + n(\theta - \bar{x})^2] \right\}. \end{aligned}$$

В этом выражении, как всегда, $\bar{x} = \sum_{i=1}^n x_i/n$, $s^2 = (\sum_{i=1}^n (x_i - \bar{x})^2)/n$ и при его выводе использовано тождественное преобразование:

$$\sum_{i=1}^n (x_i - \theta)^2 = \sum_{i=1}^n [(x_i - \bar{x}) - (\theta - \bar{x})]^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\theta - \bar{x})^2.$$

Следующий (4-й) этап, на котором вычисляется апостериорное распределение $\varphi(\theta | x_1, \dots, x_n)$, является, как правило, самым трудным в техническом отношении. Мы будем вычислять функцию $\varphi(\theta | x_1, \dots, x_n)$ с точностью до общего нормирующего множителя. Поэтому в тех местах наших выкладок, в которых мы будем опускать отдельные сомножители-компоненты этого общего нормирующего множителя, знаки равенства, связывающие наши преобразования, будут заменяться знаком эквивалентности « $\sim$ »:

$$\begin{aligned} \varphi(\theta | x_1, \dots, x_n) &\sim \exp \left\{ -\frac{1}{2} \left[\frac{(\theta - \theta_0)^2}{\sigma_0^2} + \frac{n}{\sigma^2} (s^2 + (\theta - \bar{x})^2) \right] \right\} \\ &= \exp \left\{ -\frac{1}{2} \left(\frac{\theta^2}{\sigma_0^2} - \frac{2\theta\theta_0}{\sigma_0^2} + \frac{\theta_0^2}{\sigma_0^2} + \frac{ns^2}{\sigma^2} + \frac{\theta^2 n}{\sigma^2} - \frac{2\theta n\bar{x}}{\sigma^2} + \frac{n\bar{x}^2}{\sigma^2} \right) \right\} \\ &= \exp \left\{ -\frac{1}{2} \left[\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) \theta^2 - 2 \left(\frac{\theta_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2} \right) \theta \right. \right. \\ &\quad \left. \left. + \left(\frac{\theta_0^2}{\sigma_0^2} + \frac{n\bar{x}^2}{\sigma^2} + \frac{ns^2}{\sigma^2} \right) \right] \right\} \\ &\sim \exp \left\{ -\frac{1}{2\Delta^2} (\theta - a)^2 \right\}, \end{aligned} \tag{7.36}$$

где

$$\Delta^2 = \frac{\sigma_0^2}{1 + \frac{n\sigma_0^2}{\sigma^2}} \quad \text{и} \quad a = \frac{\bar{x} + \theta_0 \frac{\sigma^2}{n\sigma_0^2}}{1 + \frac{\sigma^2}{n\sigma_0^2}}. \quad (7.37)$$

Из (7.36) следует, что апостериорное распределение $\varphi(\theta | x_1, \dots, x_n)$, аккумулирующее в себе как априорную, так и выборочную информацию о параметре θ , снова является нормальным со средним значением a и дисперсией Δ^2 , определяемыми соотношениями (7.37).

Соответственно байесовская оценка параметра θ в данном случае определяется формулой

$$\hat{\theta}_{\text{ср}}^{(\theta)} = \hat{\theta}_{\text{мод}}^{(\theta)} = \frac{\bar{x} + \theta_0 \frac{\sigma^2}{n\sigma_0^2}}{1 + \frac{\sigma^2}{n\sigma_0^2}}.$$

Как мы знаем (см. выше, п. 7.5.1), оценка параметра θ по методу максимального правдоподобия есть

$$\hat{\theta}_{\text{мп}} = \bar{x}.$$

Мы видим, что байесовская оценка получается как взвешенное среднее оценки $\hat{\theta}_{\text{мп}}$ и среднего значения θ_0 априорного распределения параметра θ , причем вес θ_0 тем больше, чем меньше дисперсия σ_0^2 априорного распределения и чем больше дисперсия анализируемой генеральной совокупности (т.е. чем менее точно производимые наблюдения x_i характеризуют среднее значение своей генеральной совокупности).

Возвращаясь к конкретным числовым данным нашего примера, имеем: $\hat{\theta}_{\text{мп}} = \bar{x} = 0,775$; $\hat{\theta}^{(\theta)} = 0,687$.

Пример 7.11. Рассмотрим задачу оценивания неизвестной вероятности θ в схеме испытаний Бернулли (см. п. 3.1.1) на основе n независимых испытаний. Таким образом, можно сказать, что мы исследуем биномиальную генеральную совокупность (см. формулу (3.3)) с неизвестным значением вероятности «успеха» θ в одном испытании, имея единственное наблюдение x_1 — число «успехов» в серии из n независимых испытаний. Распределение вероятностей $f(x | \theta)$ для числа успехов $\nu_\theta(n)$ задается формулой

$$f(x | \theta) = P\{\nu_\theta(n) = x\} = C_n^x \theta^x (1 - \theta)^{n-x}.$$

Анализ предыстории функционирования систем, аналогичных исследуемой, позволил определить в качестве *априорного распределения* $p(\theta)$

параметра θ бета-распределение с параметрами a и b (см. п. 3.2.6, формулу (3.22)):

$$p(\theta) = \begin{cases} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} & \text{при } 0 \leq \theta \leq 1; \\ 0 & \text{при } \theta \notin [0; 1]. \end{cases}$$

В нашем случае при единственном имеющемся наблюдении x_1 функция правдоподобия $L(x_1 | \theta) = f(x_1 | \theta) = C_n^{x_1} \theta^{x_1} (1-\theta)^{n-x_1}$, так что апостериорная плотность $\varphi(\theta | x_1)$ определится в соответствии с (7.34) соотношением:

$$\varphi(\theta | x_1) = \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} C_n^{x_1} \theta^{x_1} (1-\theta)^{n-x_1}}{\int_0^1 C_n^{x_1} \theta^{x_1} (1-\theta)^{n-x_1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} d\theta}.$$

Произведя очевидные сокращения в числителе и знаменателе правой части этого выражения и учитывая тот факт, что в соответствии с определениями гамма- и бета-функций

$$\begin{aligned} \int_0^1 \theta^{a-1+x_1} (1-\theta)^{b-1+n-x_1} d\theta &= B(a+x_1; b+n-x_1) \\ &= \frac{\Gamma(a+x_1) \cdot \Gamma(b+n-x_1)}{\Gamma(a+b+n)}, \end{aligned}$$

имеем:

$$\varphi(\theta | x_1) = \frac{\Gamma(a+b+n)}{\Gamma(a+x_1)\Gamma(b+n-x_1)} \theta^{a+x_1-1} (1-\theta)^{b+n-x_1-1}.$$

Но отсюда следует, что апостериорное распределение оцениваемого параметра θ снова является бета-распределением. Его среднее значение известным образом выражается через его два параметра (см. п. 3.2.6), так что

$$\hat{\theta}_{\text{ср}}^{(6)} = E(\theta | x_1) = \frac{a+x_1}{(a+x_1) + (b+n-x_1)} = \frac{x_1+a}{n+a+b} = \frac{\frac{x_1}{n} + \frac{a}{n}}{1 + \frac{a+b}{n}}.$$

Как мы знаем (см. п. 7.5.1), оценкой максимального правдоподобия параметра θ в данном примере будет отношение

$$\hat{\theta}_{\text{мп}} = \frac{x_1}{n}.$$

Сравнение $\hat{\theta}_{\text{ср}}^{(6)}$ и $\hat{\theta}_{\text{мп}}$ показывает, что эти оценки могут достаточно сильно различаться (в зависимости от значений a и b) при небольших

объемах выборок. Однако по мере роста n они, как и следует ожидать при состоятельных способах оценивания, будут давать все более схожие результаты.

Пример 7.12. Как известно (см. п. 3.1.8), длительность жизни элемента (устройства, системы, живого организма), работающего в режиме *нормальной эксплуатации*, характеризуется экспоненциальным з.р.в., т.е. функцией плотности $f(x | \theta)$ вида

$$f(x | \theta) = \begin{cases} \theta e^{-\theta x} & \text{при } x \geq 0; \\ 0 & \text{при } x < 0. \end{cases}$$

Функция правдоподобия $L(x_1, x_2, \dots, x_n | \theta)$ имеющихся в нашем распоряжении n независимых наблюдений $x_1, x_2, \dots, x_n$, соответственно, имеет вид

$$L(x_1, x_2, \dots, x_n | \theta) = \theta^n e^{-\theta \sum_{i=1}^n x_i}.$$

Максимизация по θ логарифмической функции правдоподобия

$$l(x_1, \dots, x_n | \theta) = \ln L = n \ln \theta - \theta \sum_{i=1}^n x_i$$

приводит нас к оценке *максимального правдоподобия* $\hat{\theta}_{\text{мп}}$ параметра θ вида

$$\hat{\theta}_{\text{мп}} = \frac{1}{\bar{x}(n)}, \quad \text{где } \bar{x}(n) = \frac{1}{n} \sum_{i=1}^n x_i.$$

Построим *байесовские* точечную и интервальную оценки параметра θ в предположении, что априорное распределение параметра θ подчиняется гамма-закону с параметрами a и b (см. п. 3.2.5), т.е.

$$p(\theta) = \begin{cases} \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} & \text{при } \theta \geq 0; \\ 0 & \text{при } \theta < 0. \end{cases}$$

В соответствии с формулой (7.34) *апостериорное* распределение параметра θ (с точностью до нормирующих множителей, не зависящих от θ) будет характеризоваться функцией плотности

$$\begin{aligned} \varphi(\theta | x_1, \dots, x_n) &\sim p(\theta) \cdot L(x_1, \dots, x_n | \theta) = \\ &= \begin{cases} \theta^{a+n-1} e^{-(b + \sum_{i=1}^n x_i)\theta} & \text{при } \theta \geq 0 \\ 0 & \text{при } \theta < 0. \end{cases} \end{aligned} \quad (7.38)$$

Правая часть (7.38) задает (с точностью до нормирующих множителей) снова гамма-распределение, но с параметрами $a' = a + n$ и $b' = b + \sum_{i=1}^n x_i$.

Среднее значение гамма-распределений (с параметрами a' и b') случайной величины равно, как известно (см. п. 3.2.5), отношению первого параметра ко второму, так что *точечная байесовская оценка* $\hat{\theta}^{(6)}$ параметра θ будет иметь вид:

$$\hat{\theta}^{(6)} = \frac{a + n}{b + \sum_{i=1}^n x_i} = \frac{\frac{1}{\bar{x}(n)} + \frac{a}{\sum_{i=1}^n x_i}}{1 + \frac{b}{\sum_{i=1}^n x_i}}.$$

Интервальная байесовская оценка (при уровне доверия $P_0 = 1 - 2q$) определится $100(1 - q)$ - и $100q$ -процентными точками апостериорного распределения (7.38), соответственно, $\gamma_{1-q}(a', b')$ и $\gamma_q(a', b')$, так что, с учетом известной связи между процентными точками гамма- и χ^2 -распределений ($\gamma_c(a', b') = \frac{1}{2b'} \chi_q^2(2a')$, см. (Абрамовиц М., 1979)), имеем (с вероятностью $P_0 = 1 - 2q$):

$$\frac{1}{2(b + \sum_{i=1}^n x_i)} \chi_{1-q}^2(2a + 2n) < \theta < \frac{1}{2(b + \sum_{i=1}^n x_i)} \chi_q^2(2a + 2n).$$

Пример 7.13. В данном примере речь идет об оценке параметра θ пуассоновского закона (см. п. 3.1.3, формулу (3.6)), т. е. анализируемая генеральная совокупность описывается законом распределения вероятностей вида

$$f(x | \theta) = P\{\xi = x | \theta\} = \frac{\theta^x}{x!} e^{-\theta} \quad (x = 0, 1, 2, \dots).$$

Априорное распределение вероятностей параметра θ подчиняется гамма-закону с параметрами a и b (см. п. 3.2.5, формулу (3.21)), т. е.

$$p(\theta) = \begin{cases} \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} & \text{при } \theta > 0; \\ 0 & \text{при } \theta \leq 0. \end{cases}$$

Функция правдоподобия

$$L(x_1, \dots, x_n | \theta) = \frac{1}{\prod_{i=1}^n (x_i!)} \theta^{\sum_{i=1}^n x_i} e^{-n\theta}.$$

Как известно (см. п. 7.5.1), оценка максимального правдоподобия параметра θ имеет вид

$$\hat{\theta}_{\text{мп}} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Апостериорная плотность (с точностью до нормирующего множителя) в соответствии с формулой (7.34) имеет вид

$$\varphi(\theta | x_1, \dots, x_n) \sim \theta^{a + \sum_{i=1}^n x_i - 1} e^{-(n+b)\theta},$$

который соответствует гамма-распределению с параметрами $a' = a + \sum_{i=1}^n x_i$ и $b' = n + b$. Следовательно, байесовская оценка

$$\hat{\theta}_{\text{ср}}^{(6)} = E(\theta | x_1, \dots, x_n) = \frac{a'}{b'} = \frac{a + \sum_{i=1}^n x_i}{n + b} = \frac{\hat{\theta}_{\text{мп}} + \frac{a}{n}}{1 + \frac{b}{n}}.$$

* * *

Применение байесовского подхода в эконометрике и, в частности, в статистическом оценивании неизвестных значений параметров оказывается в определенных ситуациях крайне актуальным и достаточно эффективным [А. Зельнер]. Конечно, «узким» местом в его прикладной распространенности является обоснованный выбор априорного распределения. Бывают ситуации, когда информация о виде априорного распределения имеется с точностью до небольшого числа неизвестных параметров, которые можно оценить по выборке одновременно с оцениванием параметра θ . В этом случае говорят, что применен *эмпирический байесовский подход*.

ВЫВОДЫ

1. Одна из центральных задач статистического анализа реальной системы заключается в вычислении (на основании имеющихся статистических данных) как можно более точных *приближенных значений (статистических оценок)* для одного или нескольких числовых параметров, характеризующих функционирование этой системы. Принципиальная возможность получения работоспособных приближений такого рода на основании статистического обследования *лишь части* анализируемой генеральной совокупности (т. е. на основании ограниченного ряда наблюдений, или выборки) обеспечивается замечательным свойством *статистической устойчивости* выборочных характеристик (см. п. 6.2).

2. Статистическая оценка строится в виде функции от результатов наблюдений, а потому *сама по природе является случайной величиной*.

При повторении выборки из той же самой генеральной совокупности и при подстановке новых выборочных значений в ту же самую «функцию-оценку» мы, вообще говоря, получаем другое число в качестве приближенного значения интересующего нас параметра, т. е. имеется неконтролируемый разброс в значениях оценки при повторениях эксперимента (в данном случае — выборки)!

3. В качестве основной меры точности статистической оценки $\hat{\Theta}$ неизвестного параметра Θ используется средний квадрат ее отклонения от оцениваемого значения, т. е. величина $E(\hat{\Theta} - \Theta)^2$, а в многомерном случае — ковариационная матрица компонент векторной оценки $\hat{\Theta}$. Очевидно, чем меньше эта величина (или обобщенная дисперсия оценки $\hat{\Theta}$ в многомерном случае), тем точнее (*эффективнее*) оценка. Для широкого класса генеральных совокупностей существует неравенство (неравенство Рао–Крамера–Фреше (7.12), (7.13)), задающее тот минимум $\Delta_{\min}^2$ (по всем возможным оценкам) среднего квадрата $E(\hat{\theta} - \theta)^2$, улучшить который невозможно. Естественно использовать этот минимум $\Delta_{\min}^2$ в качестве начальной точки отсчета *меры эффективности* оценки, определив эффективность $e(\hat{\theta})$ любой оценки $\hat{\theta}$ параметра θ в виде отношения

$$e(\hat{\theta}) = \frac{\Delta_{\min}^2}{E(\hat{\theta} - \theta)^2}.$$

4. Свойство *состоятельности* оценки $\hat{\theta}$ (см. п. 7.1.3) обеспечивает ее статистическую устойчивость, т. е. ее сходимость (по вероятности) к истинному значению оцениваемого параметра θ по мере роста объема выборки, на основании которой эта оценка строится. Свойство *несмещенности* оценки $\hat{\theta}$ (см. п. 7.1.4) заключается в том, что результат усреднения всевозможных значений этой оценки, полученных по различным выборкам заданного объема (из одной и той же генеральной совокупности), дает в точности истинное значение оцениваемого параметра, т. е. $E\hat{\theta} = \theta$. Далеко не всегда следует настаивать на необходимом соблюдении свойства несмещенности оценки: несущественное само по себе уже при умеренно больших объемах выборки, оно может чрезмерно обеднить класс оценок, в рамках которого решается задача построения наилучшей оценки.

5. С учетом случайной природы каждого конкретного оценочного значения $\hat{\theta}$ неизвестного параметра θ представляет интерес построение *целых интервалов* оценочных значений $\Delta\theta$, а в многомерном случае — *целых областей*, которые с наперед заданной (и близкой к единице) вероятностью P накрывали бы истинное значение оцениваемого параметра θ , т. е. $P\{\theta \in \Delta_p(\hat{\theta})\} = P$. Эти интервалы (области) принято называть *довери-*

тельными (или интервальными оценками). Существуют два подхода к построению интервальных оценок: *точный* (конструктивно реализуемый лишь в сравнительно узком классе ситуаций) и *асимптотически приближенный* (наиболее распространенный в практике статистических приложений), см. п. 7.5.4.

6. Основными методами построения статистических оценок являются:

метод максимального правдоподобия (см. п. 7.5.1);

метод моментов (см. п. 7.5.2);

метод, использующий «взвешивание» наблюдений, — цензурирование, урезание, порядковые статистики (см. п. 7.5.4).

Различные варианты метода, использующего «взвешивание» наблюдений, находят все большее распространение в связи с *устойчивостью* получаемых при этом статистических выводов по отношению к возможным отклонениям реального распределения исследуемой генеральной совокупности от постулируемого модельного.

7. Наличие априорной информации об оцениваемом параметре, позволяющей сопоставить с каждым возможным значением неизвестного параметра некую вероятностную меру его достоверности, т.е. сведений об *априорном вероятностном законе распределения оцениваемого параметра*, позволяет существенно уточнить оценки, полученные традиционными методами (методом максимального правдоподобия, методом моментов и т.п.) в условиях отсутствия такой информации. Построение таких оценок осуществляется с помощью так называемого *байесовского подхода* (см. п. 7.6), а сами оценки называются байесовскими.

Приложение к гл. 7 (доказательство неравенства информации)

Рассмотрим тождества

$$\int \dot{\theta}(x_1, \dots, x_n) L(x_1, \dots, x_n; \theta) dx_1 \dots dx_n \equiv \theta + b_{\dot{\theta}}(\theta),$$

$$\int L(x_1, \dots, x_n; \theta) dx_1 \dots dx_n \equiv 1.$$

Их почленное дифференцирование по θ (с учетом условий а) и б), см. с. 244, и обозначения $l = \ln L = \sum_{i=1}^n \ln f(x_i; \theta)$) дает

$$\int \dot{\theta} \frac{dl}{d\theta} L dx_1 \dots dx_n \equiv \mathbf{E} \left(\dot{\theta} \frac{dl}{d\theta} \right) \equiv 1 + \frac{db_{\dot{\theta}}(\theta)}{d\theta},$$

$$\int \frac{dl}{d\theta} L dx_1 \dots dx_n \equiv \mathbf{E} \left(\frac{dl}{d\theta} \right) \equiv 0, \quad \text{откуда следует :}$$

$$\mathbf{E} \left[(\dot{\theta} - \theta) \left(\frac{dl}{d\theta} - \mathbf{E} \left(\frac{dl}{d\theta} \right) \right) \right] = \mathbf{E} \left(\dot{\theta} \frac{dl}{d\theta} \right) = 1 + \frac{db_{\dot{\theta}}(\theta)}{d\theta} \quad (*)$$

Применяя к левой части (*) неравенство Коши–Буняковского ($[E(\xi\eta)]^2 \leq E\xi^2 E\eta^2$) и принимая во внимание тождества $E(dl/d\theta) = 0$, $E(dl/d\theta)^2 = n E(d \ln f(x; \theta)/d\theta)^2$, а также — равенство левой и правой частей (*), получаем

$$\left(1 + \frac{db_{\hat{\theta}}(\theta)}{d\theta}\right)^2 \leq E(\hat{\theta} - \theta)^2 n E\left(\frac{d \ln f(x; \theta)}{d\theta}\right)^2,$$

откуда и следует непосредственно неравенство информации (7.12).

ГЛАВА 8. СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ (СТАТИСТИЧЕСКИЕ КРИТЕРИИ)

На разных стадиях статистического исследования и моделирования возникает необходимость в формулировке и экспериментальной проверке некоторых предположительных утверждений (гипотез) относительно природы или величины неизвестных параметров анализируемой стохастической системы. Например, исследователь высказывает предположение: «исследуемые наблюдения извлечены из нормальной генеральной совокупности» или «среднее значение анализируемой генеральной совокупности равно нулю». Будем обозначать в дальнейшем высказанное нами предположение (гипотезу) с помощью буквы H . Наша цель — проверить, не противоречит ли высказанная нами гипотеза H имеющимся выборочным данным.

Процедура обоснованного сопоставления высказанной гипотезы с имеющимися в нашем распоряжении выборочными данными $x_1, x_2, \dots, x_n$, сопровождаемая количественной оценкой степени достоверности получаемого вывода, осуществляется с помощью того или иного *статистического критерия* и называется *статистической проверкой гипотез*.

Результат подобного сопоставления может быть либо *отрицательным* (данные наблюдения противоречат высказанной гипотезе, а потому от этой гипотезы следует отказаться), либо *неотрицательным* (данные наблюдения не противоречат высказанной гипотезе, а потому ее можно принять в качестве одного из естественных и допустимых решений). При этом неотрицательный результат статистической проверки гипотезы не означает, что высказанное нами предположительное утверждение является наилучшим, единственно подходящим: просто она не противоречит имеющимся у нас выборочным данным, однако таким же свойством могут наряду с H обладать и другие гипотезы. Так что даже статистически проверенное предположение H следует расценивать не как раз и навсегда установленный, абсолютно верный факт, а лишь как

достаточно правдоподобное, не противоречащее опыту утверждение.

По своему прикладному содержанию высказываемые в ходе статистической обработки данных гипотезы можно подразделить на несколько основных типов.

8.1. Основные типы гипотез, проверяемых в ходе статистического анализа и моделирования

8.1.1. Гипотезы о типе закона распределения исследуемой случайной величины

При обработке ряда наблюдений

$$x_1, x_2, \dots, x_n \quad (8.1)$$

исследуемой случайной величины ξ очень важно понять механизм формирования выборочных значений x_i , т. е. подобрать и обосновать некоторую модельную функцию распределения $F_{\text{мод}}(x)$ (например, из числа описанных в гл. 3), с помощью которой можно адекватно описать исследуемую функцию распределения $F_{\xi}(x)$. На определенной стадии исследования это приводит к необходимости проверки гипотез типа

$$H: F_{\xi}(x) \equiv F_{\text{мод}}(x), \quad (8.2)$$

где гипотетичная модельная функция может быть как заданной *однозначно* (тогда $F_{\xi}(x) = F_0(x)$, где $F_0(x)$ — полностью известная функция), так и заданной *с точностью до принадлежности к некоторому параметрическому семейству* (тогда $F_{\text{мод}}(x) = F(x; \Theta)$, где Θ — некоторый, вообще говоря, k -мерный параметр, значения которого неизвестны, но могут быть оценены по выборке (8.1) с помощью методов, изложенных в п. 7.5 и 7.6).

Проверка гипотез типа (8.2) осуществляется с помощью так называемых *критериев согласия* и опирается на ту или иную меру различия между анализируемой эмпирической функцией распределения $\hat{F}_{\xi}^{(n)}(x)$ и гипотетическим модельным законом $F_{\text{мод}}(x)$ (примеры реализации подобных критериев см. в п. 8.6.1).

8.1.2. Гипотезы об однородности двух или нескольких обрабатываемых выборок или некоторых характеристик анализируемых совокупностей

Наиболее типичные задачи такого рода характеризуются следующей общей ситуацией. Пусть мы имеем несколько «порций» выборочных данных типа (8.1):

$$\begin{aligned} 1 - \text{я: } & x_{11}, x_{12}, \dots, x_{1n_1}; \\ 2 - \text{я: } & x_{21}, x_{22}, \dots, x_{2n_2}; \\ & \dots\dots\dots \\ l - \text{я: } & x_{l1}, x_{l2}, \dots, x_{ln_l}. \end{aligned} \quad (8.3)$$

Эти порции могли образоваться, например, естественным образом — в ходе проведения выборочного обследования (скажем, за счет раздельности условий их регистрации во времени или пространстве). Обозначая функцию распределения, описывающую вероятностный закон, которому подчиняются наблюдения j -й выборки, с помощью $F_j(x)$ и снабжая тем же индексом все интересующие нас эмпирические и теоретические характеристики этого закона (средние значения $\bar{a}_j$ и a_j ; дисперсии $\hat{\sigma}_j^2$ и σ_j^2 и т. д.), основные гипотезы однородности можно записать в виде:

$$H_F: F_1(x) \equiv F_2(x) \equiv \dots \equiv F_l(x); \quad (8.4a)$$

$$H_a: a_1 = a_2 = \dots = a_l; \quad (8.4б)$$

$$H_\sigma: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_l^2. \quad (8.4в)$$

В случае неотрицательного результата проверки этих гипотез говорят, что соответствующие выборочные характеристики (например, $\hat{a}_1, \hat{a}_2, \dots, \hat{a}_l$) различаются *статистически незначимо*.

Отметим частный случай гипотез типа (8.4а), когда число выборок $l = 2$, а одна из выборок содержит малое количество наблюдений (в частном случае — одно). В таком виде проверка гипотез типа (8.4а) означает *проверку аномальности одного или нескольких резко выделяющихся наблюдений*. Реализация конкретных критериев однородности описана в п. 8.6.2.

8.1.3. Гипотезы о числовых значениях параметров исследуемой генеральной совокупности

Пусть, например, ряд наблюдений (8.1) дает нам значения некоторого параметра изделий, измеренные на n изделиях, случайно отобранных из массовой продукции определенного станка автоматической линии, и пусть

a_0 — заданное номинальное значение этого параметра. Каждое отдельное значение x_i может, естественно, как-то отклоняться от заданного номинала. Очевидно, для того чтобы проверить правильность настройки этого станка, надо убедиться в том, что среднее значение параметра у производимых на нем изделий будет соответствовать номиналу, т.е. проверить гипотезу типа

$$H: E\xi = a_0. \quad (8.5)$$

В общем случае гипотезы подобного типа имеют вид:

$$H_0: \Theta \in \Delta_0, \quad (8.6)$$

где Θ — некоторый параметр (вообще говоря, многомерный), от которого зависит исследуемое распределение, а Δ_0 — область его конкретных гипотетических значений, которая может состоять всего из одной точки.

Статистическая проверка гипотез о числовых значениях параметров играет важную роль в эконометрическом моделировании, регрессионном анализе, в широком спектре задач статистического исследования зависимостей, существующих между анализируемыми показателями (этому кругу вопросов посвящены гл. 10, 11 и второй том данного учебника). В частности, принятие решения о включении или исключении той или иной переменной в анализируемую регрессионную (эконометрическую) модель, о наличии-отсутствии статистической связи между наблюдаемыми признаками существенно опирается обычно на проверку гипотез типа (8.6) при $\Delta_0 = 0$. Такого же типа гипотезы приходится проверять при установлении факта независимости и стационарности имеющегося ряда наблюдений (см. п. 3.1 тома 2).

Реализация некоторых конкретных критериев статистической проверки гипотез о числовых значениях параметров анализируемой стохастической системы описывается в п. 8.6.3.

8.1.4. Гипотезы об общем виде модели, описывающей статистическую зависимость между признаками

В п. 8.1.1 речь шла, по существу о *подборе подходящей модели* для описания закона распределения вероятностей исследуемой случайной величины. Не менее важное место в общем статистическом и эконометрическом анализе занимает проблема подбора подходящей модели, с помощью которой мы можем адекватно описать исследуемую статистическую зависимость между анализируемыми признаками. В качестве гипотетических могут проверяться утверждения о линейном, квадратическом, экспоненциальном, степенном, логарифмическом, полиномиальном и т.п. типе ис-

комой зависимости. Подробнее эта проблематика обсуждается в гл. 10 данного учебника.

8.2. Общая логическая схема статистического критерия

По своему назначению и характеру решаемых задач статистические критерии чрезвычайно разнообразны. Однако их объединяет общность логической схемы, по которой они строятся. Коротко эту логическую схему можно описать так.

1. Выдвигается гипотеза H_0 .

2. Задаются величиной так называемого *уровня значимости* критерия α . Дело в том, что всякое статистическое решение, т. е. решение, принимаемое на основании *ограниченного* ряда наблюдений, неизбежно сопровождается некоторой, хотя, возможно, может и очень малой, вероятностью ошибочного заключения как в ту, так и в другую сторону. Скажем, в какой-то небольшой доле случаев α гипотеза H_0 может оказаться отвергнутой, в то время как на самом деле она является справедливой, или, наоборот, в какой-то небольшой доле случаев β мы можем принять нашу гипотезу, в то время как на самом деле она ошибочна, а справедливым оказывается некоторое конкурирующее с ней предположение — альтернативная гипотеза H_1 . При фиксированном объеме выборочных данных величину вероятности *одной из этих ошибок* мы можем выбирать по своему усмотрению. Если же объем выборки можно как угодно увеличивать, то имеется принципиальная возможность добиваться как угодно малых вероятностей обеих ошибок α и β при любом фиксированном конкурирующем предположительном утверждении H_1 . В частности, при фиксированном объеме выборки обычно задаются величиной α вероятности ошибочного отвержения проверяемой гипотезы H_0 , которую часто называют «основной» или «нулевой». Эту вероятность ошибочного отклонения «нулевой» гипотезы принято называть *уровнем значимости* или *размером критерия*. Выбор величины уровня значимости α зависит от сопоставления потерь, которые мы понесем в случае ошибочных заключений в ту или иную сторону: чем весомее для нас потери от ошибочного отвержения высказанной гипотезы H_0 , тем меньшей выбирается величина α . Однако поскольку такое сопоставление в большинстве практических задач оказывается весьма затруднительным (часто трудно даже вообще сказать, в какую сторону ошибка является для нас более опасной), то, как правило, пользуются некоторыми стандартными значениями уровня значимости. К таким стандартным значениям можно причислить величины

$\alpha = 0,1; 0,05; 0,025; 0,01; 0,005; 0,001$. Особенно распространенной является величина уровня значимости α , равная 0,05. Она означает, что в среднем в пяти случаях из 100 мы будем ошибочно отвергать высказанную гипотезу при многократном использовании данного статистического критерия.

3. Задаются некоторой функцией от результатов наблюдения (*критической статистикой*) $\gamma^{(n)} = \gamma(x_1, x_2, \dots, x_n)$. Эта критическая статистика $\gamma^{(n)}$, как и всякая функция от результатов наблюдения, сама является случайной величиной (см. 7.1.2) и в предположении справедливости гипотезы H_0 подчинена некоторому хорошо изученному (затабулированному) закону распределения с плотностью $f_{\gamma^{(n)}}(u)$.

Один из основных принципов построения критической статистики (*принцип отношения правдоподобия*) описан в п. 8.3.1. Поясним здесь лишь общий содержательный смысл этой статистики: как правило, ею определяется мера расхождения имеющихся в нашем распоряжении выборочных данных (8.1) с высказанной (и проверяемой) гипотезой H_0 . Так, в гипотезах типа рассмотренных в п. 8.1.1 критическая статистика $\gamma^{(n)}$ определяет меру различия между анализируемой эмпирической функцией распределения $\hat{F}^{(n)}(x)$ и гипотетической (модельной) функцией $F_{\text{мод}}(x)$. В гипотезах типа рассмотренных в п. 8.1.2 величина $\gamma^{(n)}$ измеряет степень расхождения соответствующих выборочных характеристик в различных выборках; в гипотезах типа рассмотренных в п. 8.1.3 — отклонения выборочных характеристик от соответствующих гипотетических значений и т. д.

4. Из таблиц распределения $f_{\gamma^{(n)}}(u)$ находятся $100(1 - \frac{\alpha}{2})\%$ -ная точка $\gamma_{\alpha/2}^{\min}$ и $100\frac{\alpha}{2}\%$ -ная точка $\gamma_{\alpha/2}^{(\max)}$, разделяющие всю область мыслимых значений случайной величины $\gamma^{(n)}$ на три части: *область неправдоподобно малых* (I), *неправдоподобно больших* (III) и естественных или *правдоподобных* (в условиях справедливости гипотезы H_0) значений (II) (рис. 8.1). В тех случаях, когда основную опасность для нашего утверждения представляют только односторонние отклонения, т. е. только «слишком маленькие» или только «слишком большие» значения критической статистики $\gamma^{(n)}$, находят лишь одну процентную точку: либо $100(1 - \alpha)\%$ -ную точку $\gamma_{\alpha}^{(\min)}$, которая будет разделять весь диапазон значений $\gamma^{(n)}$ на две части: область неправдоподобно малых и область правдоподобных значений; либо $100\alpha\%$ -ную точку $\gamma_{\alpha}^{(\max)}$; она будет разделять весь диапазон значений $\gamma^{(n)}$ на область неправдоподобно больших и область правдоподобных значений.

5. Наконец, в функцию $\gamma^{(n)}$ подставляют имеющиеся конкретные вы-

борочные данные $x_1, \dots, x_n$ и подсчитывают численную величину $\gamma^{(n)}$. Если окажется, что вычисленное значение принадлежит области правдоподобных значений $\gamma^{(n)}$, то гипотеза H_0 считается не противоречащей выборочным данным. В противном случае, т.е. если $\gamma^{(n)}$ слишком мала или слишком велика, делается вывод, что $\gamma^{(n)}$ на самом деле не подчиняется закону $f_{\gamma^{(n)}}(u)$ (этот вывод, как легко понять, сопровождается вероятностью ошибки, равной α), и это несоответствие мы вынуждены объяснить ошибочностью высказанного нами предположения H_0 и, следовательно, отказаться от него.

Таким образом, решение, принимаемое на основании любого статистического критерия, может оказаться ошибочным как в случае отклонения проверяемой гипотезы H_0 (с вероятностью α), так и в случае ее принятия (с вероятностью β). Вероятности α и β ошибочных решений называют также *ошибками* соответственно *первого* и *второго рода*, а величину $1 - \beta$ — *мощностью критерия*. Очевидно, из двух критериев, характеризующихся одной и той же вероятностью α отвергнуть в действительности правильную гипотезу H_0 , следует предпочесть тот, который сопровождается меньшей ошибкой второго рода (или большей мощностью).

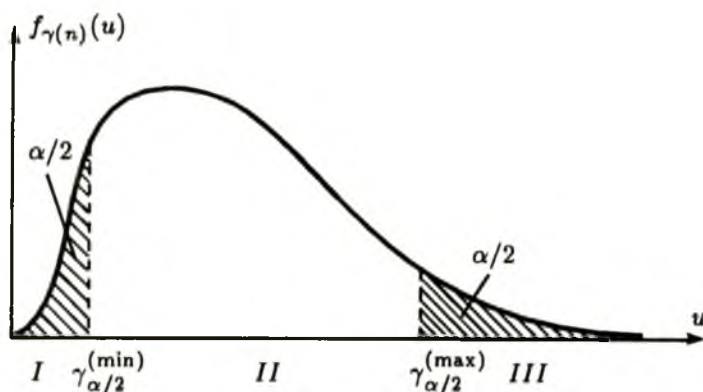


Рис. 8.1. График плотности распределения критической статистики γ_n и выделение областей «правдоподобных» (II) и «неправдоподобных» (I и III), в условиях справедливости гипотезы H_0 , значений этой статистики

Если проверяемое предположительное утверждение сводится к гипотезе о том, что значение некоторого параметра Θ в *точности* равно заданной величине Θ_0 (см. выше гипотезы, рассмотренные в п. 8.1.3), то эта гипотеза называется *простой*. В других случаях гипотеза будет на-

ываться сложной.

8.3. Построение статистического критерия; принцип отношения правдоподобия

Попытаемся выяснить, как конкретно получаются те функции от результатов наблюдения (критические статистики $\gamma^{(n)}$), по значениям которых принимается окончательное решение о том, соответствует ли проверяемая гипотеза имеющимся у нас данным (8.1) или противоречит им.

8.3.1. Сущность принципа отношения правдоподобия

Для пояснения общего принципа, приводящего к построению наилучших (наиболее мощных при заданной величине уровня значимости) критериев, вернемся к условиям примера с заработной платой (см. п. 7.5.1) и соответственно к рис. 7.2. В этом примере исследовалась логарифмически нормально распределенная заработная плата ξ работников определенной совокупности, а в качестве исходных данных мы располагали тремя наблюдениями (тремя обследованными работниками): $x_1 = 190$ ден.ед.; $x_2 = 175$ ден.ед.; $x_3 = 205$ ден.ед. Пусть мы хотим проверить основную гипотезу (простую) о среднем значении нормально распределенной случайной величины $\ln \xi$:

$$H_0: E(\ln \xi) = 5,240 = a_0$$

против простой альтернативы

$$H_1: E(\ln \xi) = 5,443 = a_1.$$

Из рис. 7.2 видно, что гипотеза H_0 не противоречит имеющимся наблюдениям (более того, в данном случае наши наблюдения выглядят наиболее правдоподобными именно при гипотезе H_0), в то время как те же наблюдения оказываются маловероятными в условиях справедливости гипотезы H_1 .

В общем случае представление о сравнительной правдоподобности имеющихся наблюдений $x_1, \dots, x_n$ (в отношении проверяемой и альтернативной гипотез) дает нам сопоставление соответствующих функций правдоподобия (см. формулу (7.5)) и, в частности, их отношение

$$\gamma^{(n)} = \frac{L_{H_1}(x_1, \dots, x_n; \Theta)}{L_{H_0}(x_1, \dots, x_n; \Theta)} = \frac{L(x_1, \dots, x_n; \Theta_1)}{L(x_1, \dots, x_n; \Theta_0)}, \quad (8.7)$$

где L_{H_1} и L_{H_0} — значения функций правдоподобия наблюдений $x_1, \dots, x_n$, вычисленные в предположении справедливости соответственно гипотез $H_1: \Theta = \Theta_1$ и $H_0: \Theta = \Theta_0$. Очевидно, чем правдоподобнее наблюдения в условиях гипотезы H_0 , тем больше функция правдоподобия L_{H_0} и тем меньше величина $\gamma^{(n)}$. Если $f_{\gamma^{(n)}}(u)$ — плотность распределения статистики $\gamma^{(n)}$ при условии справедливости гипотезы H_0 , то построение критерия проверки гипотезы H_0 с заданным уровнем значимости α сводится к определению 100 α %-ной точки γ_α распределения $f_{\gamma^{(n)}}(u)$ и к реализации следующего правила:

$$\left. \begin{array}{l} \text{если } \gamma^{(n)} > \gamma_\alpha, \text{ то гипотеза } H_0 \text{ отвергается} \\ \text{с вероятностью ошибиться, равной } \alpha, \text{ так как} \\ \text{в соответствии с законом } f_{\gamma^{(n)}}(u) \text{ и при справед-} \\ \text{ливости гипотезы } H_0 \text{ возможно осуществление} \\ \text{события } \{\gamma^{(n)} > \gamma_\alpha\} \text{ с вероятностью } \alpha, \text{ т.е.} \\ \int\limits_{\gamma_\alpha}^{\infty} f_{\gamma^{(n)}}(u) du = \alpha; \\ \text{если } \gamma^{(n)} < \gamma_\alpha, \text{ то гипотеза } H_0 \text{ не отвергается.} \end{array} \right\} \quad (8.8)$$

Критерии, основанные на статистиках $\gamma^{(n)}$ вида (8.7) и процедурах (8.8), носят название *критериев отношения правдоподобия*, а их практическая реализуемость и предпочтительность по отношению к другим возможным критериям подкреплены следующими фактами (справедливыми в достаточно широком классе ситуаций).

1. Критерии отношения правдоподобия являются наиболее мощными среди всех других возможных критериев (этот факт для случая сравнения двух простых гипотез сформулирован и доказан в виде леммы Неймана–Пирсона, см., например [Крамер Г.]).

2. Плотность $f_{\gamma^{(n)}}(u)$ распределения критической статистики $\gamma^{(n)}$, как правило, без труда восстанавливается по функции правдоподобия L наблюдаемой случайной величины.

Так, в рассмотренном выше примере с проверкой гипотезы о среднем значении нормальной случайной величины ξ (при известном значении

дисперсии σ^2), имеем:

$$\begin{aligned} L_{H_j} &= L(x_1, \dots, x_n; \Theta_j) = L(x_1, \dots, x_n; a_j, \sigma^2) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - a_j)^2} \quad (j = 0, 1), \end{aligned}$$

так что

$$\begin{aligned} \gamma^{(n)} &= \frac{L_{H_1}}{L_{H_0}} = e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n [(x_i - a_1)^2 - (x_i - a_0)^2]} \\ &= e^{c \frac{(x - a_0)\sqrt{n}}{\sigma} - \frac{1}{2} c^2}, \end{aligned}$$

где $c = \sqrt{n} (a_1 - a_0) \cdot \frac{1}{\sigma}$.

Поскольку в нашем примере $n = 3$; $\sigma^2 = 0,16$; $a_1 = 5,443$ и $a_0 = 5,240$, то $c > 0$, и если мы положим

$$\gamma_\alpha = e^{c u_{Q_\alpha} - \frac{1}{2} c^2},$$

где $Q_\alpha = 100 \cdot 2\alpha\%$ и u_{Q_α} — $Q_\alpha\%$ -ная точка стандартного нормального распределения, то неравенство

$$e^{c \frac{(x - a_0)\sqrt{n}}{\sigma} - \frac{1}{2} c^2} > \gamma_\alpha$$

будет выполняться на множестве всех таких выборок $(x_1, \dots, x_n)$, для которых $\sqrt{n}(x - a_0) \frac{1}{\sigma} > u_{Q_\alpha}$, или, что то же,

$$\bar{x} \geq a_0 + u_{Q_\alpha} \cdot \frac{\sigma}{\sqrt{n}}.$$

Получившееся правило проверки гипотезы *не зависит от альтернативного значения параметра a_1* , а потому является (принимая во внимание лемму Неймана–Пирсона) наиболее мощным *при всех возможных* альтернативных значениях параметра $a_1 > a_0$, или, как принято в таких случаях говорить, *равномерно наиболее мощным*.

8.3.2. Критерий логарифма отношения правдоподобия для проверки простой гипотезы

Пусть известно, что ряд наблюдений $x_1, \dots, x_n$ можно рассматривать как независимую выборку из распределения, принадлежащего семейству распределений $F(x; \Theta)$, где Θ — k -мерный параметр. Требуется проверить гипотезу о том, что $\Theta = \Theta_0$ (гипотеза (8.6), п. 8.1.3).

Рассмотрим критерий

$$\tilde{\gamma}^{(n)} = -2 \ln \{L(x_1, \dots, x_n; \Theta_0) / L(x_1, \dots, x_n; \tilde{\Theta})\}. \quad (8.9)$$

где $\tilde{\Theta}$ — оценка (введенная в п. 7.5.1) максимального правдоподобия (ОМП) параметра Θ по выборке $x_1, \dots, x_n$. Доказано, что при наложении на семейство $F(x; \Theta)$ и на значение Θ_0 дополнительных требований, гарантирующих оптимальные свойства оценок максимального правдоподобия (см. п. 7.5.1), величина $\tilde{\gamma}^{(n)}$ имеет асимптотически (при $n \rightarrow \infty$) χ^2 -распределение с k степенями свободы (см. [Пао С. Р.]).

В качестве примера применения критерия $\tilde{\gamma}^{(n)}$ рассмотрим еще раз задачу проверки гипотезы о среднем значении нормальной совокупности, приведенную в п. 8.3.1. В введенных там обозначениях с учетом того, что оценка максимального правдоподобия параметра a_0 есть $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, имеем:

$$\tilde{\gamma}^{(n)} = 2 \left(\sum_{i=1}^n \frac{(x_i - a_0)^2}{2\sigma^2} - \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{2\sigma^2} \right) = n(\bar{x} - a_0)^2 / \sigma^2.$$

Поскольку в предположении справедливости гипотезы H_0 статистика $\bar{x}$ нормально распределена со средним a_0 и дисперсией σ^2/n , $\tilde{\gamma}^{(n)}$ имеет χ^2 -распределение с одной степенью свободы (в данном случае это результат *точный*, а не *асимптотический*!).

8.3.3. Критерий отношения правдоподобия для проверки сложной гипотезы

Рассмотрим модификацию критерия (8.9) для случая, когда в гипотезе конкретизируются значения не всех параметров, как в предыдущем пункте, а лишь части из них. Пусть $\Theta = (\theta_1, \dots, \theta_k)$ — вектор неизвестных параметров распределения и гипотеза состоит в том, что

$$H_{01}: \theta_j = \theta_{0j}, \quad j = 1, \dots, r < k. \quad (8.10)$$

Удобно разбить вектор Θ на две части: $\Theta_1 = (\theta_1, \dots, \theta_r)$ и $\Theta_2 = (\theta_{r+1}, \dots, \theta_k)$. Обозначим $\tilde{\Theta}_2$ оценку максимального правдоподобия Θ_2 по выборке $x_1, \dots, x_n$ при известном значении $\Theta_1 = \Theta_{01}$ и $(\hat{\Theta}_1, \hat{\Theta}_2)$ — оценку максимального правдоподобия (Θ_1, Θ_2) . Критерий для проверки гипотезы H_{01} имеет вид:

$$\tilde{\gamma}^{(n)} = -2 \ln \{L(x_1, \dots, x_n; \Theta_{01}, \tilde{\Theta}_2) / L(x_1, \dots, x_n; \hat{\Theta}_1, \hat{\Theta}_2)\}. \quad (8.11)$$

Можно доказать, что при выполнении ряда дополнительных требований, гарантирующих оптимальные свойства оценок максимального правдоподобия (см. п. 7.5.1), величина $\hat{\gamma}^{(n)}$ имеет асимптотически (при $n \rightarrow \infty$) χ^2 -распределение с r степенями свободы.

8.4. Характеристики качества статистического критерия

Характеристиками точности статистического критерия проверки простых или сложных гипотез типа (8.6) служат:

$\alpha(\Theta)$ — вероятность отвергнуть основную гипотезу H_0 , подсчитанная в предположении, что истинное значение «проверяемого» параметра равно Θ ; величину $1 - \alpha(\Theta)$ называют *оперативной характеристикой* критерия, а значение $\alpha(\Theta_0)$ в задаче проверки простой гипотезы « $H_0: \Theta = \Theta_0$ » есть не что иное, как уровень значимости (размер) критерия (вероятность ошибки первого рода);

$\beta(\Theta)$ — вероятность отвергнуть конкурирующую (с основной) гипотезу, подсчитанная в предположении, что истинное значение «проверяемого» параметра равно Θ ; величину $1 - \beta(\Theta)$ называют *мощностью* (или функцией мощности) критерия, а значение $\beta(\Theta_1)$ в задаче проверки простой гипотезы « $H_0: \Theta = \Theta_0$ » против альтернативы « $H_1: \Theta = \Theta_1$ » есть не что иное, как вероятность ошибки второго рода.

Обсудим эти характеристики и зависящие от них свойства критерия.

Пусть Γ_n — область возможных значений критической статистики $\gamma^{(n)}$, а $\Gamma_n^{H_0}$ и $\Gamma_n^{H_1}$ — описанные в п. 8.2 соответственно области «правдоподобных» и «неправдоподобных» (в условиях справедливости гипотезы H_0) значений $\gamma^{(n)}$. Тогда очевидно:

$$\left. \begin{aligned} \alpha_n(\Theta) &= P_{\Theta} \{ \gamma^{(n)} \in \Gamma_n^{H_1} \} = \int_{\Gamma_n^{H_1}} f_{\gamma^{(n)}}(u, \Theta) du; \\ \beta_n(\Theta) &= 1 - \alpha_n(\Theta) = P_{\Theta} \{ \gamma^{(n)} \in \Gamma_n^{H_0} \} \\ &= \int_{\Gamma_n^{H_0}} f_{\gamma^{(n)}}(u; \Theta) du \end{aligned} \right\}, \quad (8.12)$$

где $P_{\Theta}\{A\}$ и $f_{\gamma^{(n)}}(u; \Theta)$ — соответственно вероятность события A и плотность распределения критической статистики $\gamma^{(n)}$, подсчитанные в предположении, что истинное значение проверяемого параметра равно Θ .

В условиях проверки параметрических гипотез вида (8.6) с заданным уровнем значимости α_0 критерий $\{\gamma^{(n)}, \Gamma_n^{H_0}\}$ называется *несмещенным*, если

$$\begin{cases} \alpha_n(\Theta) \leq \alpha_0 & \text{при всех } \Theta \in \Delta_0; \\ \alpha_n(\Theta) > \alpha_0 & \text{при всех } \Theta \notin \Delta_0. \end{cases}$$

И, наконец, критерий $\{\gamma^{(n)}, \Gamma_n^{H_0}\}$ называется *состоятельным*, если

$$\lim_{n \rightarrow \infty} \alpha_n(\Theta) = 1 \quad \text{при всех } \Theta \notin \Delta_0.$$

Последнее соотношение означает, в частности, что функция мощности состоятельного критерия $1 - \beta_n(\Theta)$ стремится (при $n \rightarrow \infty$) к единице при любом значении Θ , не входящем в область Δ_0 гипотетичных (в соответствии с гипотезой H_0) значений параметра.

Из (8.12) очевидно, что при любом фиксированном объеме выборки n перестройка критерия в направлении уменьшения уровня значимости α (т. е. сужения области $\Gamma_n^{H_1}$) связана с одновременным увеличением ошибки второго рода, а в общем случае — с уменьшением значений функции мощности $1 - \beta(\Theta)$ (так как при этом расширяется область $\Gamma_n^{H_0}$ отклонения альтернативы H_1). И наоборот, перестройка критерия (в любом фиксированном классе критериев, в том числе и в классе наиболее мощных критериев) в направлении увеличения его мощности связана (при фиксированном объеме выборки n) с неизбежным одновременным увеличением его уровня значимости. В то же время неограниченным увеличением объема выборки (т. е. при $n \rightarrow \infty$) можно добиваться сколь угодно малых значений для вероятностей ошибок вида

$$\begin{aligned} \alpha_n^* &= \begin{cases} \alpha_n(\Theta^*), & \text{где } \Theta^* \in \Delta_0^* \subset \Delta_0, \text{ или} \\ \max_{\Theta \in \Delta_0} \alpha_n(\Theta); \end{cases} \\ \beta_n^* &= \begin{cases} \beta_n(\Theta^{**}), & \text{где } \Theta^{**} \notin \Delta_0, \text{ или} \\ \max_{\Theta \notin \Delta_0} \beta_n(\Theta). \end{cases} \end{aligned}$$

Для больших объемов выборок (т. е. асимптотически по $n \rightarrow \infty$) существуют соотношения, связывающие между собой характеристики α_n^*, β_n^* и n (см., например [Кендалл М. Дж., Стьюарт А., 1973. С. 310–311]). Остановимся здесь на одном полезном соотношении такого типа, позволяющем, в частности, определять объем выборки $n(\alpha, \beta; p)$, необходимый в критерии отношения правдоподобия (Неймана–Пирсона) для различения двух простых гипотез

H_0 : выборка извлечена из генеральной совокупности

$$f(x; \Theta_0); \quad (8.13)$$

H_1 : выборка извлечена из генеральной совокупности

$$f(x; \Theta_1)$$

с ошибками первого и второго рода, не превосходящими заданных значений соответственно α и β ; величина $\rho = \rho(H_0, H_1)$, от которой также зависит объем выборки n , характеризует «расстояние» между гипотезами H_0 и H_1 и определяется по формуле $\rho(H_0, H_1) = \int \ln \frac{f(x; \Theta_1)}{f(x; \Theta_0)} [f(x; \Theta_1) - f(x; \Theta_0)] dx$, где интегрирование ведется по всей области возможных значений наблюдаемой случайной величины x , а $f(x; \Theta)$ — ее плотность распределения. В [Айвазян С. А., 1959] показано, что в достаточно широком классе случаев при различении близких простых гипотез (т. е. при малых значениях ρ) справедлива приближенная (асимптотическая) формула

$$n(\alpha; \beta; \rho) \approx \frac{(u_{1-\alpha} + u_{1-\beta})^2}{\rho(H_0, H_1)}, \quad (8.14)$$

в которой u_q , как и прежде, квантиль уровня q (или q -квантиль) стандартного нормального закона распределения.

З а м е ч а н и е. Обратим внимание читателя на практическую неизбежность проявления двух «невыгодных» для всей теории проверки статистических гипотез эффектов: *эффекта «слишком малого объема выборки»* и *эффекта «слишком большого объема выборки»*.

Эффект «слишком малого объема выборки» состоит в том, что при заданной величине уровня значимости критерия α и малом числе наблюдений n , на основании которых принимается решение, мощность критерия, т. е. вероятность отклонить проверяемую «нулевую» гипотезу H_0 в ситуации, когда она в действительности не имеет места, оказывается слишком маленькой (приближенное представление о взаимосвязи величин α , β и n дает формула (8.14)). Есть два выхода из этой ситуации: либо увеличить объем выборки n , либо несколько увеличить уровень значимости α , что повлечет соответствующее уменьшение β (т. е. увеличение мощности критерия $1 - \beta$).

Для пояснения эффекта «слишком большого объема выборки» подчеркнем следующее: никто в действительности не считает, что какая-либо гипотеза выполняется точно: мы просто строим абстрактную модель реальных событий, которая в какой-то мере обязательно отклоняется от истины; однако, как мы видим, огромная выборка почти наверняка (т. е. с

вероятностью, стремящейся к единице при неограниченном возрастании n)¹ отвергает в этом случае нашу гипотезу при любом заданном уровне значимости α .

Казалось бы, налицо «тупиковая» ситуация: при малых выборках вывод статистически ненадежен, а при слишком больших — однозначно предопределен. Так, например, авторы неоднократно наблюдали обескураживающее действие эффекта больших n на прикладника-исследователя, пытающегося с помощью критериев согласия подобрать подходящий модельный закон $F_{\text{мод}}(x)$ для описания распределения исследуемой генеральной совокупности и неизменно приходящего при этом к отрицательному результату (т.е. к отвержению проверяемой гипотезы).

Чтобы избежать эффекта большой выборки, априорное задание характеристик точности критерия (уровня значимости α и ошибки второго рода β) необходимо увязывать с объемом имеющихся данных n : выигрыш в «чувствительности» критерия, получающийся в результате увеличения n , целесообразно использовать для уменьшения как α , так и β . В частности, если определить уменьшение α при возрастании n , то очень малые отклонения от H_0 уже не приведут к обязательному отвержению этой гипотезы: вероятность этого факта будет зависеть от того, с какой «скоростью» (с ростом n) убывает α .

8.5. Последовательная схема принятия решения (последовательные критерии)

8.5.1. Последовательная схема наблюдений

Если число наблюдений, на основании которых статистик принимает решение, не фиксируется заранее, но ставится в зависимость от результатов зарегистрированных на каждой данной стадии эксперимента наблюдений, то говорят об использовании *последовательной схемы наблюдений*. Поскольку результаты наблюдений на каждой фиксированной стадии эксперимента представляют собой случайную выборку из генеральной совокупности и, следовательно, случайны по своей природе, то и момент прекращения наблюдений (определение которого зависит от этих результатов) также является величиной случайной.

Впервые идея об использовании последовательной схемы наблюдений возникла в ходе конструирования экономных планов выборочного статисти-

¹ Это следует из свойства состоятельности критерия (см. выше).

стического контроля качества продукции¹. Речь шла о выборочной проверке того факта, что неизвестная нам истинная доля p дефектных изделий во всем производстве не превосходит некоторого порогового значения (предельно допустимой доли брака) p_0 . Авторы упомянутой работы предложили для этой цели методику проверки с двукратной выборкой. На первом этапе предложенной ими последовательной схемы извлекается одна выборка изделий объема n_1 и оценивается доля брака $\hat{p}_1$ в этой выборке. Решение о необходимости контроля второй выборки принимается на основании результатов наблюдений в первой выборке: грубо говоря, если доля брака в первой выборке $\hat{p}_1$ оказалась существенно меньше (или существенно больше) заданного порогового значения p_0 , то необходимости в повторной выборке нет и принимается гипотеза $p \leq p_0$ (или соответственно альтернатива $p > p_0$); если же доля брака $\hat{p}_1$ в первой выборке несущественно отличается от заданного порогового значения p_0 (или, как говорят, находится в «зоне неопределенности», или в «зоне безразличия»), то принимается решение о продолжении наблюдений и, в частности, об извлечении второй выборки объема n_2 .

При такой методике достигается заметный выигрыш (в среднем) в числе наблюдений, необходимом для различения интересующих нас гипотез с заданными характеристиками точности (уровнем значимости α и мощностью $1 - \beta$) по сравнению с критерием Неймана–Пирсона, являющимся наилучшим (наиболее мощным, см. п. 8.3.1) среди всех критериев, основанных на классической схеме наблюдений (т. е. построенных на базе выборок *заранее заданного объема* n). Поэтому к последовательной схеме наблюдений целесообразно обращаться в ситуациях, когда каждое наблюдение является дорогостоящим или труднодоступным и по условиям эксперимента исследователь имеет практическую возможность реализовать эту схему (далеко не всегда исследователь находится в подобных условиях).

К величинам $\alpha(\Theta)$ и $\beta(\Theta)$, характеризующим «качество» всякого критерия (см. п. 8.4), при рассмотрении последовательных критериев добавляется еще *средний объем выборки* $E_{\Theta} n(\alpha, \beta)$, необходимый для проверки гипотез вида (8.6) с заданными характеристиками точности (α, β) (нижний индекс Θ при знаке математического ожидания E означает, что усреднение проводится при истинном значении параметра, равном Θ).

¹ Dodge H. F., Romig H. G. A method of sampling inspection — The Bell System Techn. Journ., 8 (1929), pp. 613–631.

8.5.2. Последовательный критерий отношения правдоподобия (критерий Вальда) и его свойства

Построение статистического критерия при *фиксированном объеме выборки* n (см. п. 8.3.1) сводится в конечном счете к разбиению области возможных значений критической статистики $\gamma^{(n)} = \gamma(x_1, \dots, x_n)$ на две части: область правдоподобных и область неправдоподобных (в условиях справедливости проверяемой гипотезы H_0) значений $\gamma^{(n)}$. При попадании конкретного значения $\gamma(x_1, \dots, x_n)$ в область неправдоподобных значений принимается решение об отклонении проверяемой гипотезы.

Последовательный критерий, т.е. критерий, основанный на последовательной схеме наблюдений, построен по той же логической схеме с одним отличием: последовательно для каждого фиксированного объема выборки $\nu = 1, 2, \dots, n, n+1, \dots$ область Γ_ν возможных значений критической статистики $\gamma(x_1, \dots, x_\nu)$ разбивается на три непересекающихся части: область $\Gamma_\nu^{H_0}$ правдоподобных, область Γ_ν^H неправдоподобных и область Γ_ν^* сомнительных (в условиях справедливости проверяемой гипотезы H_0) значений, т.е.

$$\Gamma_\nu = \Gamma_\nu^{H_0} \cup \Gamma_\nu^H \cup \Gamma_\nu^*, \quad \nu = 1, 2, \dots$$

На каждом ν -м шаге последовательной схемы наблюдений, т.е. при наличии наблюдений $x_1, \dots, x_\nu$, $\nu = 1, 2, \dots$, решение принимается по следующему правилу:

если $\gamma(x_1, \dots, x_\nu) \in \Gamma_\nu^{H_0}$, то проверяемая гипотеза H_0 принимается;

если $\gamma(x_1, \dots, x_\nu) \in \Gamma_\nu^H$, то проверяемая гипотеза H_0 отвергается (или принимается некоторая альтернатива H_1);

если $\gamma(x_1, \dots, x_\nu) \in \Gamma_\nu^*$, то окончательный вывод откладывается и производится следующее $(\nu+1)$ -е наблюдение (поэтому область Γ_ν^* иногда называют областью неопределенности или областью продолжения наблюдений).

Таким образом, для того чтобы иметь какой-то конкретный статистический критерий, надо конкретизировать: а) тип проверяемой гипотезы; б) способ построения критической статистики $\gamma(x_1, \dots, x_\nu)$; в) способ построения областей $\Gamma_\nu^{H_0}$, Γ_ν^H и Γ_ν^* по заданным (требуемым) значениям характеристик точности критерия.

В качестве конкретного примера последовательного критерия рассмотрим известный критерий отношения правдоподобия Вальда, предназначенный для различения двух простых гипотез вида (8.13).

Критическая статистика этого критерия для последовательности не-

зависимых наблюдений $x_1, \dots, x_\nu$ определяется соотношением

$$\gamma^{(\nu)} = \ln \frac{f(x_1; \Theta_1) \dots f(x_\nu; \Theta_1)}{f(x_1; \Theta_0) \dots f(x_\nu; \Theta_0)} = \sum_{i=1}^{\nu} \ln \frac{f(x_i; \Theta_1)}{f(x_i; \Theta_0)}, \quad \nu = 1, 2, \dots \quad (8.15)$$

Области правдоподобных ($\Gamma_\nu^{H_0}$), неправдоподобных ($\Gamma_\nu^{H_1}$) и сомнительных (Γ_ν^*), в условиях справедливости гипотезы H_0 , значений критической статистики $\gamma^{(\nu)}$ приближенно задаются соотношениями:

$$\begin{aligned} \Gamma_\nu^{H_0} &= \left\{ \gamma: \gamma^{(\nu)} \leq \ln \frac{\beta}{1-\alpha} \right\}; \\ \Gamma_\nu^{H_1} &= \left\{ \gamma: \gamma^{(\nu)} \geq \ln \frac{1-\beta}{\alpha} \right\}; \\ \Gamma_\nu^* &= \left\{ \gamma: \ln \frac{\beta}{1-\alpha} < \gamma^{(\nu)} < \ln \frac{1-\beta}{\alpha} \right\}. \end{aligned} \quad (8.16)$$

Была доказана *оптимальность* этого критерия среди всех других возможных последовательных критериев, а именно: среди всех критериев, различающих гипотезы (8.13) с ошибками первого и второго рода, не превосходящими заданных величин, соответственно α и β , *критерий* (8.15)–(8.16) *требует наименьшего среднего числа наблюдений* $E_{\Theta_i} \nu(\alpha, \beta)$ как в условиях справедливости гипотезы $H_0 (i=0)$, так и в условиях справедливости гипотезы $H_1 (i=1)$ [Вальд].

Исследования показали, что этот критерий примерно в два–четыре раза выгоднее (по затратам на наблюдения), чем наилучший из классических критериев — критерий Неймана–Пирсона [Айвазян С. А., 1959].

8.6. Методы проверки статистических гипотез: примеры статистических критериев

8.6.1. Критерии согласия

Критерии согласия предназначены для статистической проверки гипотез о модельном виде закона распределения вероятностей (з.р.в.) исследуемой случайной величины. Пусть $F_{\text{мод}}(x; \theta_1, \dots, \theta_k)$ — гипотетичный модельный вид функции распределения анализируемого з.р.в. Функция $F_{\text{мод}}(x; \theta_1, \dots, \theta_k)$ может *однозначно* задавать исследуемый з.р.в. (тогда все значения параметров $\theta_1, \dots, \theta_k$ известны априори), а может обозначать

лишь гипотетичный тип искомого з.р.в., *определяемый целым параметрическим семейством* $\{F_{\text{мод}}(x; \theta_1, \dots, \theta_k)\}$ (например, мы хотим проверить гипотезу о том, что з.р.в. анализируемой случайной величины описывается нормальной моделью, см. п. 3.1.5).

Таким образом, описываемые в данном пункте критерии согласия предназначены для проверки гипотезы

$$H_0: F_{\xi}(x) \in \{F_{\text{мод}}(x; \theta_1, \dots, \theta_k)\}, \quad (8.17)$$

на основании n независимых наблюдений (8.1) анализируемой случайной величины ξ , функция распределения которой, как обычно, обозначена с помощью $F_{\xi}(x)$. Критические статистики этих критериев основаны на различных мерах расстояний между анализируемой эмпирической функцией распределения $\hat{F}^{(n)}(x)$ (см. п. 6.2.1) и гипотетическим семейством $\{F_{\text{мод}}(x; \theta_1, \dots, \theta_k)\}$.

Там, где речь идет о *группированных* выборочных данных, мы будем придерживаться терминологии и обозначений п. 6.1.

Критерий согласия χ^2 Пирсона. Этот критерий позволяет проверять гипотезы вида (8.17) как для дискретных, так и для непрерывных случайных величин в ситуациях, когда параметры $\theta_1, \dots, \theta_k$ априори известны или являются неизвестными. Метод проверки (критерий) основан на теореме Пирсона–Фишера, которые доказали, что *если гипотеза* (8.17) *истинна*, то при некоторых достаточно общих условиях распределение критической статистики

$$\gamma^{(n)} = \sum_{j=1}^s \frac{(\nu_j - np_j(\hat{\Theta}))^2}{np_j(\hat{\Theta})} \quad (8.18)$$

сходится (при $n \rightarrow \infty$) к $\chi^2(s - k - 1)$ -распределению.

Для объяснения обозначений в соотношении (8.18) разберем отдельно случаи непрерывной и дискретной случайных величин.

В *непрерывном* случае статистика (8.18) строится по *группированным* данным. Группировка данных (8.1) производится в соответствии с рекомендациями и обозначениями п. 6.1, т. е.: s — это общее число интервалов группирования; ν_j — число выборочных данных, попавших в j -й интервал группирования; $\Theta = (\theta_1, \dots, \theta_k)$ — векторный параметр, который участвует в выражении модельной функции распределения $F_{\text{мод}}(x; \theta_1, \dots, \theta_k)$, а $\hat{\Theta}$ — его состоятельная оценка (более корректным способом оценивания считается тот, при котором в качестве $\hat{\Theta}$ используется оценка максимального правдоподобия, построенная по группированным данным); при этом

k в теореме Пирсона–Фишера обозначает число неизвестных параметров, *оцениваемых по выборке*, так что в случае априори известных значений параметров $k = 0$; наконец, $p_j(\hat{\Theta})$ — это результат модельного расчета вероятности попасть в j -й интервал группирования, т. е.

$$p_j(\hat{\Theta}) = F_{\text{мод}}(c_j; \hat{\Theta}) - F_{\text{мод}}(c_{j-1}; \hat{\Theta}), \quad (8.19)$$

где c_{j-1}, c_j — соответственно левый и правый концы j -го интервала группирования ($j = 1, 2, \dots, s$).

Случай анализа *дискретной* случайной величины отличается от предыдущего лишь тем, что мы работаем с *исходной* (а не группированной) выборкой (8.1), ν_j — это число выборочных данных, равных j -му возможному значению x_j^0 , s — число всех возможных значений случайной величины, а

$$p_j(\hat{\Theta}) = P\{\xi = x_j^0\} = F_{\text{мод}}(x_{j+1}^0) - F_{\text{мод}}(x_j^0). \quad (8.19')$$

Итак, теорема Пирсона–Фишера дает нам основание следующим образом построить процедуру проверки гипотезы (8.17).

1) По выборочным данным (8.1) (сгруппированным при анализе непрерывной случайной величины) строим состоятельные оценки $\hat{\Theta}$ параметров $\Theta = (\theta_1, \dots, \theta_k)$.

2) В соответствии с формулами (8.18)–(8.19)–(8.19') подсчитываем значение критической статистики $\gamma^{(n)}$.

3) По заданному уровню значимости критерия α из таблиц процентных точек χ^2 -распределения находим $100(1 - \alpha/2)$ - и $100\alpha/2\%$ -ные точки $\chi_{1-\alpha/2}^2(s - k - 1)$ и $\chi_{\alpha/2}^2(s - k - 1)$ «хи-квадрат» распределения с $s - k - 1$ степенями свободы.

4) Если $\chi_{1-\alpha/2}^2(s - k - 1) < \gamma^{(n)} < \chi_{\alpha/2}^2(s - k - 1)$, то гипотеза (8.17) не отклоняется; если же $\gamma^{(n)} \leq \chi_{1-\alpha/2}^2(s - k - 1)$ или $\gamma^{(n)} \geq \chi_{\alpha/2}^2(s - k - 1)$, то гипотеза (8.17) отклоняется¹.

¹ Отвержение выдвинутой гипотезы в случае «слишком маленьких» значений статистики критерия $\gamma^{(n)}$ на первый взгляд противоречит здравому смыслу. Действительно, статистика $\gamma^{(n)}$ характеризует степень отклонения эмпирического распределения случайной величины ξ от гипотетического F : чем меньше $\gamma^{(n)}$, тем меньше это отклонение. Казалось бы, идеальный случай достигается при $\gamma^{(n)} \equiv 0$. Однако надо отметить, что хотя $\gamma^{(n)}$ и является мерой отклонения гипотетического закона F от истинного, но мерой случайной, т. е. величиной, подверженной обязательному неконтролируемому разбросу. И в этом отношении одинаково неправдоподобными (маловероятными) следует считать как слишком большие значения $\gamma^{(n)}$, так и слиш-

Так, например, при проверке гипотезы нормальности гипотетический закон $F_{\text{мод}}$ будет иметь вид:

$$F_{\text{мод}}(x; a, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(u-a)^2}{2\sigma^2}} du,$$

а в качестве оценок $\hat{a}$ и $\hat{\sigma}^2$ двух неизвестных параметров a и σ^2 будут фигурировать величины

$$\hat{a} = \frac{1}{n} \sum_{j=1}^s \nu_j x_j^0;$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{j=1}^s \nu_j (x_j^0 - \hat{a})^2$$

(через x_j^0 обозначается, как и прежде, средняя точка интервала Δ_j).

Значения $F(c_j; \hat{a}, \hat{\sigma}^2) = \Phi(c_j; \hat{a}, \hat{\sigma}^2)$, необходимые для подсчета вероятностей p_i , можно найти, например, из таблицы значений функции стандартного нормального распределения с учетом соотношения $\Phi(x; a, \sigma^2) = \Phi(\frac{x-a}{\sigma}; 0, 1)$. Число степеней свободы закона распределения χ^2 , процентные точки которого нам понадобятся, будет равно в данном случае $s - 3$, где s — число интервалов группирования.

Пример 8.1. В качестве анализируемой случайной величины η рассматривается среднедушевой месячный доход (в условных денежных единицах) домашнего хозяйства некоторого региона. Обследовано 750 домашних хозяйств этого региона ($n = 750$). В таблицах 8.1а и 8.1б приведены результаты группировки выборочных данных (y_i) и их логарифмов ($x_i = \ln y_i$) соответственно (ширина интервала группирования и с х о д н ы х данных равна 25 усл. ден. ед.).

ком малые. О чем же свидетельствуют слишком малые значения $\gamma^{(n)}$? Одной из причин подобной «ненормальности» может служить неудачный выбор закона F — искусственное завышение числа параметров, от которых этот закон зависит. Другими причинами могут служить нарушения корректности или объективности техники выборочного обследования, стремление «подогнать» экспериментальные данные под желаемый результат и т. п.

Таблица 8.1а

Номер интервала (j)	Число выборочных данных, попавших в этот интервал (ν_j)	Номер интервала (j)	Число выборочных данных, попавших в этот интервал (ν_j)
1	2	21	1
2	15	22	3
3	44	23	1
4	83	24	2
5	108	25	0
6	110	26	1
7	83	27	0
8	75	28	1
9	49	29	1
10	34	30	2
11	27	31	0
12	21	32	1
13	24	33	0
14	13	34	0
15	13	35	0
16	19	36	0
17	8	37	1
18	3	38	0
19	2	39	0
20	2	40	1

Таблица 8.16

Номер интервала (j)	Число выборочных данных, попавших в этот интервал (ν'_j)	Номер интервала (j)	Число выборочных данных, попавших в этот интервал (ν'_j)
1	1	10	141
2	1	11	123
3	0	12	86
4	7	13	65
5	8	14	52
6	15	15	14
7	39	16	8
8	73	17	5
9	109	18	3

Воспользуемся критерием согласия χ^2 Пирсона для проверки нормальности случайной величины $\xi = \ln \eta$ по данным, представленным в табл. 8.16 (одновременно, соответственно, проверяется *логарифмическая* нормальность данных, представленных в табл. 8.1а).

Оценим вначале неизвестные значения параметров $a = E\xi$ и $\sigma^2 = D\xi$ гипотетического нормального распределения по группированным данным, представленным в табл. 8.16:

$$\hat{a} = \frac{1}{750} \sum_{j=1}^{18} \nu'_j x_j^0 = 4,99 \approx 5,00$$

$$\hat{\sigma}^2 = \frac{1}{749} \sum_{j=1}^{18} \nu'_j \cdot (x_j^0 - \hat{a})^2 = 0,366; \quad \hat{\sigma} = 0,605.$$

В данном случае

$$p_j(\hat{\Theta}) = p_j(\hat{a}; \hat{\sigma}^2) = \Phi\left(\frac{c_j - \hat{a}}{\hat{\sigma}}\right) - \Phi\left(\frac{c_{j-1} - \hat{a}}{\hat{\sigma}}\right),$$

где c_j — правый конец j -го интервала группирования ($j = 1, \dots, 18$), $c_0 = 2,5$, а $\Phi(z)$ — значение функции распределения стандартного нормального закона в точке z (определяется из табл. П1.2, см. Приложение 1).

Результаты остальных расчетов приведены в табл. 8.2.

Таблица 8.2

Номер интер- вала (j)	c_j	$\frac{c_j - \hat{a}}{\hat{\sigma}}$	$\Phi\left(\frac{c_j - \hat{a}}{\hat{\sigma}}\right)$	$p_j(\hat{\Theta}) =$ $= \Phi\left(\frac{c_j - \hat{a}}{\hat{\sigma}}\right) -$ $-\Phi\left(\frac{c_{j-1} - \hat{a}}{\hat{\sigma}}\right)$	$np_j(\hat{\Theta})$	$\frac{(\nu'_j - np_j(\hat{\Theta}))^2}{np_j(\hat{\Theta})}$
0	2,50	-4,13	$10^{-6} \cdot 18$	—	—	—
1	2,75	-3,72	$10^{-6} \cdot 100$	$82 \cdot 10^{-6}$	0,062	14,19
2	3,00	-3,30	$10^{-6} \cdot 484$	$3,84 \cdot 10^{-4}$	0,288	1,76
3	3,25	-2,89	$10^{-6} \cdot 1926$	$1,44 \cdot 10^{-3}$	1,082	1,08
4	3,50	-2,48	$10^{-6} \cdot 6569$	$4,64 \cdot 10^{-3}$	3,484	3,55
5	3,75	-2,07	$10^{-6} \cdot 19226$	$12,66 \cdot 10^{-3}$	9,493	0,23
6	4,00	-1,65	$10^{-6} \cdot 49472$	$30,25 \cdot 10^{-3}$	22,684	2,60
7	4,25	-1,24	0,107488	0,058016	43,512	0,47
8	4,50	-0,83	0,203269	0,095781	71,836	0,02
9	4,75	-0,41	0,340903	0,137634	103,226	0,32
10	5,00	0,00	0,500	0,159097	119,323	3,94
11	5,25	0,41	0,659	0,159	119,250	0,18
12	5,50	0,83	0,797	0,138	103,500	2,96
13	5,75	1,24	0,893	0,096	72,000	0,68
14	6,00	1,65	0,951	0,058	43,500	1,66
15	6,25	2,07	0,981	0,030	22,500	3,21
16	6,50	2,48	0,993	0,012	9,000	0,11
17	6,75	2,89	0,998	0,005	3,750	0,42
18	7,00	3,30	0,999	0,001	0,750	6,75
Σ						44,13

Проверка гипотезы о нормальности анализируемых наблюдений основана на сравнении значения критической статистики (8.18) с табличными значениями процентных точек $\chi^2(s-k-1)$ -распределения (см. табл. П1.4). В нашем случае число интервалов группирования $s = 18$, число оцененных по выборке параметров $k = 2$. Задавшись уровнем значимости критерия $\alpha = 0,05$, имеем:

$$\chi^2_{1-\frac{\alpha}{2}}(s-k-1) = \chi^2_{0,975}(15) = 6,26; \quad \chi^2_{\frac{\alpha}{2}}(s-k-1) = \chi^2_{0,025}(15) = 27,49,$$

так что, поскольку $\gamma^{(n)} = 44,13 > 27,49$, мы вынуждены отклонить гипотезу о нормальности анализируемой случайной величины.

Заметим, правда, что решающий вклад в завышенное значение кри-

тической статистики $\gamma^{(n)}$ внесли слишком *тяжелые «хвосты»* эмпирического распределения (см. слагаемые этой статистики, соответствующие первому и последнему интервалам группирования). Поэтому изображенные на рис. 8.2а и 8.2б гистограммы и модельные (соответственно, логнормальная и нормальная) плотности анализируемого распределения демонстрируют вполне приличное соответствие между выборочными и гипотетическими модельными законами.

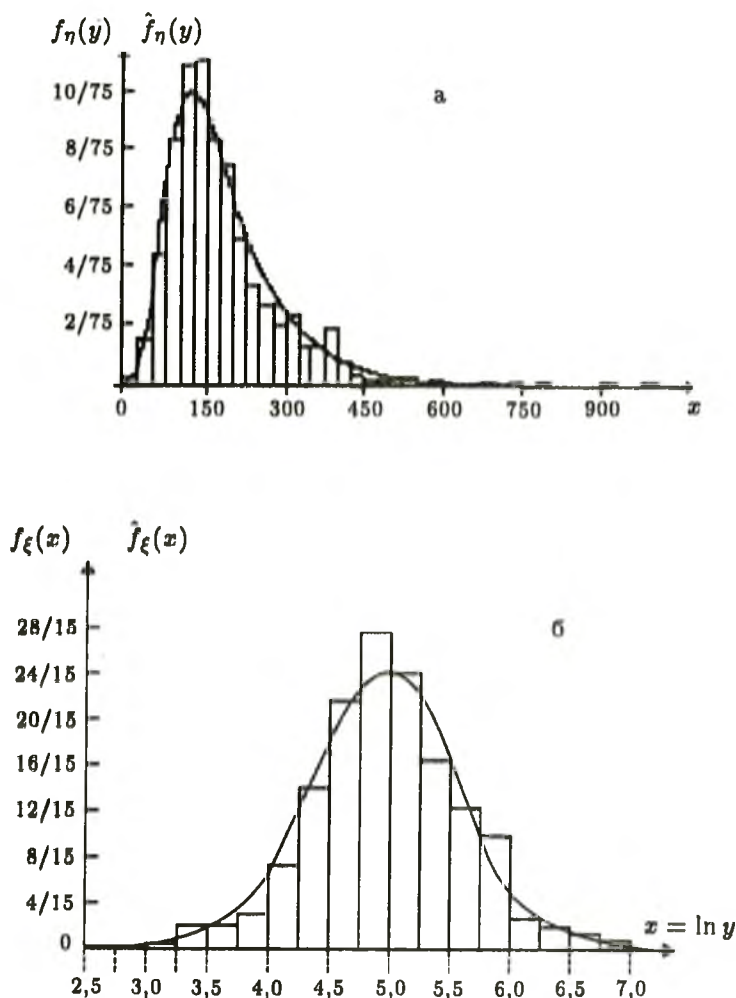


Рис. 8.2. Гистограмма и теоретическая (модельная) плотность, характеризующие распределение семей по среднему месячному доходу (а) и по логарифму среднего месячного дохода (б)

Критерий Колмогорова–Смирнова. Этот критерий позволяет осуществлять проверку гипотезы (8.17) в условиях, когда модельная функция $F_{\text{мод}}(x) = F_0(x)$ известна полностью, т. е. не зависит от неизвестных параметров.

Пусть $\hat{F}^{(n)}(x)$ — эмпирическая функция распределения. Введем следующие меры уклонения (расстояния) между функциями $\hat{F}^{(n)}(x)$ и $F_0(x)$:

$$D_n = \sup_x |F^{(n)}(x) - F_0(x)|;$$

$$D_n^+ = \sup_x (F^{(n)}(x) - F_0(x));$$

$$D_n^- = \sup_x (F_0(x) - F^{(n)}(x)).$$

Статистики $\sqrt{n}D_n^+$ и $\sqrt{n}D_n^-$ являются статистиками критериев Колмогорова и Смирнова соответственно. При этом

$$D_n = \max(D_n^+, D_n^-).$$

Для практического использования критерия Колмогорова–Смирнова статистики D_n, D_n^+, D_n^- представляются в виде:

$$D_n^+ = \max_{1 \leq i \leq n} \left(\frac{i}{n} - t_i \right);$$

$$D_n^- = \max_{1 \leq i \leq n} \left(t_i - \frac{i-1}{n} \right);$$

$$D_n = \max_{1 \leq i \leq n} (D_n^+, D_n^-),$$

где $t_i = F_0(x_{(i)})$, т. е. это значение гипотетической функции распределения, взятой в i -й точке вариационного ряда.

Для статистик D_n, D_n^+, D_n^- известны точные распределения [Большев Л. Н., Смирнов Н. В.]. Здесь приведем лишь распределение для D_n^+

$$P\{D_n^+ \leq d\} = \begin{cases} 0, & d \leq 0; \\ 1 - \sum_{j=0}^{[n-d]} C_n^j \left(\frac{j}{n} + d \right)^{j-1} \\ \quad \times \left(1 - \frac{j}{n} - d \right)^{n-j}, & 0 < d < 1; \\ 1, & d \geq 1, \end{cases}$$

где $[n - nd]$ — целая часть числа $n - nd$.

Для практических целей обычно достаточно *предельных* распределений статистик

$$\lim_{n \rightarrow \infty} P(\sqrt{n}D_n < \lambda) = K(\lambda);$$

$$\lim_{n \rightarrow \infty} P(\sqrt{n}D_n^+ < \lambda) = S(\lambda),$$

где

$$K(\lambda) = \begin{cases} 1 - 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 \lambda^2}, & \lambda > 0; \\ 0 & \lambda < 0; \end{cases}$$

$$S(\lambda) = \begin{cases} 1 - 2^{-2\lambda^2}, & \lambda \geq 0; \\ 0 & \lambda < 0. \end{cases}$$

Предельное распределение для статистики $\sqrt{n}D_n^-$ в точности совпадает с $S(\lambda)$.

8.6.2. Критерии однородности

Переходим к описанию некоторых наиболее распространенных статистических критериев, предназначенных для проверки гипотез об однородности двух или нескольких анализируемых генеральных совокупностей, где однородность понимается в том или другом смысле (см. выше, п. 8.1.2).

Критерий однородности Смирнова. Этот критерий предназначен для проверки гипотезы совпадения з.р.в. в двух или нескольких генеральных совокупностях по группированным выборкам, извлеченным из этих совокупностей, т. е. речь идет о проверке гипотез типа (8.4а). Пусть имеется l выборок (8.3) объемов соответственно $n_1, n_2, \dots, n_l$ и пусть данные каждой из этих выборок сгруппированы так, как это описано в п. 6.1, причем разбиение диапазонов исследуемых случайных величин на интервалы группирования во всех выборках произведено *одинаковым способом* (при этом выбор общего диапазона варьирования анализируемого признака во всех выборках определяется наименьшим, по всем выборкам, из минимальных выборочных значений и наибольшим из максимальных выборочных значений). Итак, мы имеем s одних и тех же для всех выборок интервалов группирования и пусть ν_{ij} — количество элементов i -й выборки, попавших в j -й интервал ($i = 1, 2, \dots, l$; $j = 1, 2, \dots, s$). В качестве

критической статистики критерия используется величина

$$\gamma^{(n)} = n \sum_{j=1}^s \sum_{i=1}^l \frac{(\nu_{ij} - \frac{\nu_{i.}\nu_{.j}}{n})^2}{\nu_{i.}\nu_{.j}} = n \sum_{j=1}^s \sum_{i=1}^l \frac{\nu_{ij}^2}{\nu_{i.}\nu_{.j}} - 1, \quad (8.20)$$

где $\nu_{i.} = \sum_{j=1}^s \nu_{ij} = n_i$ — общее число элементов i -й выборки, $\nu_{.j} = \sum_{i=1}^l \nu_{ij}$ — общее (по всем выборкам) число выборочных данных, попавших в j -й интервал группирования и $n = \sum_{i=1}^l \sum_{j=1}^s \nu_{ij}$ — суммарный (по всем выборкам) объем выборочных данных.

Н. В. Смирновым было доказано, что при неограниченном росте объемов всех выборок *и в условиях справедливости проверяемой гипотезы* (8.4а) з.р.в. статистики (8.20) стремится к закону χ^2 с числом степеней свободы, равным $(l-1)(s-1)$. Поэтому, в соответствии с общей логикой статистического критерия (см. выше, п. 8.2) гипотеза (8.4а) отвергается, если $\gamma^{(n)} \leq \chi_{1-\frac{\alpha}{2}}^2((l-1)(s-1))$ или $\gamma^{(n)} \geq \chi_{\frac{\alpha}{2}}^2((l-1)(s-1))$, и принимается при всех остальных значениях критической статистики $\gamma^{(n)}$ (здесь, как обычно, $\chi_q^2(m)$ — табличное значение 100 q %-ной точки «хи-квадрат» распределения с m степенями свободы).

В частном случае *двух* выборок (т.е. при $l = 2$) статистика (8.20) может быть записана в виде

$$\gamma^{(n)} = n_1 n_2 \sum_{j=1}^s \frac{\left(\frac{\nu_{1j}}{n_1} - \frac{\nu_{2j}}{n_2}\right)^2}{\nu_{1j} + \nu_{2j}}, \quad (8.20')$$

и при условии справедливости гипотезы однородности она будет приблизительно распределена (при больших объемах n_1 и n_2) по закону χ^2 с $s-1$ степенью свободы.

Пример 8.2. В табл. 8.2 приведены условные данные о заработной плате работников двух отраслей. В данном примере число сравниваемых генеральных совокупностей $l = 2$ и объемы выборок $n_1 = n_2 = 100$. Проверим (при уровне значимости критерия $\alpha = 0,05$) гипотезу о том, что распределения работников по заработной плате не отличаются друг от друга в двух анализируемых отраслях.

Вычисление критической статистики (8.20') в данном случае дает:

$$\gamma^{(n)} = n_1 n_2 \sum_{j=1}^8 \frac{\left(\frac{\nu_{1j}}{n_1} - \frac{\nu_{2j}}{n_2}\right)^2}{\nu_{1j} + \nu_{2j}} = \sum_{j=1}^8 \frac{(\nu_{1j} - \nu_{2j})^2}{\nu_{1j} + \nu_{2j}} = 14,58.$$

Из таблиц процентных точек χ^2 -распределения (см. приложение 1) определяем пятипроцентную точку «хи-квадрат»-распределения с семью степенями свободы $\chi^2_{0,05}(7) = 14,07$. Следовательно, гипотезу о совпадении вероятностных распределений работников по заработной плате в двух анализируемых отраслях мы должны отвергнуть ($\gamma^{(n)} > \chi^2_{0,05}(7)$). Вероятность ошибиться при этом равна 0,05.

Таблица 8.3

Номер интерв. группы (j)	Интервал заработной платы (усл. ден. ед.)	Кол-во работников, попавших в данный интервал		$\nu_{1j} + \nu_{2j}$	$\nu_{1j} - \nu_{2j}$
		(ν_{1j}) из выборки 1-й отрасли	(ν_{2j}) из выборки 2-й отрасли		
1	130–150	4	1	5	3
2	150–170	4	1	5	3
3	170–200	15	8	23	7
4	200–250	51	43	94	8
5	250–300	22	34	56	-12
6	300–350	3	7	10	-4
7	350–400	1	3	4	-2
8	400–500	-	3	3	-3

Описанный метод проверки однородности относится к *непараметрическим критериям*, так как используемая в нем критическая статистика никак не зависит от наших предположений относительно параметрического общего вида анализируемых распределений (или, как иногда говорят, «свободна от распределения». В этом его преимущество перед параметрическими критериями. Однако его реализация требует достаточно больших объемов анализируемых выборок (по меньшей мере они должны содержать по несколько десятков наблюдений). Следующий критерий также относится к непараметрическим, однако требования к объемам выборок здесь существенно мягче.

Критерий Вилкоксона–Манна–Уитни. Этот критерий предназначен для проверки однородности двух генеральных совокупностей, понимаемой в смысле отсутствия различий в значениях параметров местоположений (средних значений, медиан) соответствующих распределений (но не тождественного совпадения распределений, как это было в предыдущем критерии). Т. е. речь идет о проверке гипотез типа (8.46).

Итак, мы располагаем выборками (8.3), извлеченными из двух генеральных совокупностей ($l = 2$). Пронумеруем эти выборки так, чтобы обеспечить выполнение неравенства $n_1 \leq n_2$. Объединим выборки и по

объединенной выборке объема $n_1 + n_2$ построим общий вариационный ряд. Обозначим $R_i^{(1)}$ — порядковый номер (*ранг*), который получает при этом i -й член вариационного ряда, *построенного только по первой выборке*, в общем вариационном ряду (т. е. $R_i^{(1)}$ — это ранг элемента $x_{1(i)}$ в *общем* вариационном ряду, где $x_{1(1)}, x_{1(2)}, \dots, x_{1(n_1)}$ — вариационный ряд, построенный только по первой выборке).

Критическая статистика описываемого критерия имеет вид

$$\gamma^{(n)} = \sum_{i=1}^{n_1} R_i^{(1)} \quad (8.21)$$

и носит название «*суммы рангов*».

Удалось показать, что в условиях справедливости проверяемой гипотезы (8.46) статистика (8.21) ведет себя (асимптотически по $n_1 \rightarrow \infty$ так, что $\lim(n_1/n_2) = c > 0$) как нормально распределенная случайная величина с параметрами

$$\begin{aligned} a &= E\gamma^{(n)} = \frac{1}{2}n_1(n_1 + n_2 + 1), \\ \sigma^2 &= D\gamma^{(n)} = \frac{1}{12}n_1n_2(n_1 + n_2). \end{aligned} \quad (8.22)$$

При этом сходимость к нормальному распределению очень быстрая: оно уже эффективно работает при $n_1 > 8$.

Это позволяет построить следующее правило проверки гипотезы (8.46).

1) По заданному уровню значимости критерия α с помощью таблиц квантилей (процентных точек) стандартного нормального распределения определяем квантиль $u_{1-\alpha/2}$ уровня $1-\alpha/2$ (или $100\alpha/2\%$ -ную точку) стандартного нормального распределения.

2) Вычисляем стандартизованное значение критической статистики $\gamma^{(n)}$

$$\gamma_{\text{ст.}}^{(n)} = \frac{\gamma^{(n)} - \frac{1}{2}n_1(n_1 + n_2 + 1)}{\sqrt{\frac{1}{12}n_1n_2(n_1 + n_2)}}, \quad (8.23)$$

где значение $\gamma^{(n)}$ вычислено по формуле (8.21).

3) Если окажется, что

$$|\gamma_{\text{ст.}}^{(n)}| > u_{1-\alpha/2},$$

то проверяемую гипотезу следует отвергнуть (и соответственно принять при всех других значениях стандартизованной критической статистики $\gamma_{\text{ст.}}^{(n)}$).

Отметим в заключение, что существуют таблицы процентных точек критической статистики (8.21) (подсчитанные при условии справедливости гипотезы однородности) и *для доасимптотического случая*, т. е. для $n \leq 8$ (см. [Большев Л. Н., Смирнов Н. В.]).

Строго говоря, описанные ниже критерии однородности (критерий Стьюдента, или t -критерий; критерий дисперсионного отношения, критерий Бартлетта и др.) применимы только к выборкам (8.3), извлеченным из *нормальной* генеральной совокупности: в этом случае, как легко понять, неотрицательный результат одновременной проверки однородности средних значений (т. е. гипотезы (8.4б)) и дисперсий (т. е. гипотезы (8.4в)) достаточен для неотрицательного вывода по поводу гипотезы об однородности *самих законов распределения* (т. е. гипотезы (8.4а)). Специальные исследования показали, однако, что t -критерий является (особенно при больших объемах выборок n) весьма *устойчивым* по отношению к отклонениям исследуемых генеральных совокупностей от нормальных. А это значит, что *он может применяться и к выборкам из негауссовских генеральных совокупностей* с той лишь оговоркой, что истинные значения уровня значимости и мощности критерия в этом случае будут несколько отличаться от заданных.

Критерий Стьюдента (t -критерий). Этот критерий предназначен для проверки гипотезы однородности средних значений (8.4б) в *двух нормальных* генеральных совокупностях, *имеющих одинаковую* (хотя и неизвестную) *дисперсию* σ^2 . В качестве *критической статистики* в данном критерии используется величина

$$\gamma^{(n)} = \frac{\bar{x}_1(n_1) - \bar{x}_2(n_2)}{\bar{s} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad (8.24)$$

где $\bar{x}_j(n_j)$ и $s_j^2(n_j)$ — соответственно выборочные средние и выборочные дисперсии, построенные по j -й выборке ($j = 1, 2$), а $\bar{s}^2$ вычисляется по выборочным дисперсиям $s_1^2(n_1)$ и $s_2^2(n_2)$ по формуле

$$\bar{s}^2 = \frac{1}{n_1 + n_2 - 2} [n_1 s_1^2(n_1) + n_2 s_2^2(n_2)]. \quad (8.25)$$

При анализе статистики нормального закона (см. п. 6.2.8, формулу (6.37')) нами было показано, что в условиях справедливости гипотезы (8.4б) (и при дополнительном условии равенства дисперсий в двух анализируемых генеральных совокупностях) статистика (8.24) должна подчиняться распределению Стьюдента с $n_1 + n_2 - 2$ степенями свободы. Поэтому, определив из таблиц (при заданном уровне значимости критерия α)

100 α /2%-ную точку $t_{\alpha/2}(n_1 + n_2 - 2)$ t -распределения с $n_1 + n_2 - 2$ степенями свободы, мы *принимая решение об отклонении гипотезы однородности* (8.46), *если окажется, что $\gamma^{(n)}$, вычисленная по формулам (8.24)–(8.25), по абсолютной величине превзойдет значение $t_{\alpha/2}(n_1 + n_2 - 2)$.*

З а м е ч а н и е. Слишком большое значение статистики (8.24), т. е. такое, при котором отвергается проверяемая гипотеза однородности, может быть следствием как *статистически значимого расхождения выборочных средних* (т. е. невыполнения гипотезы (8.46)), так и *статистически значимого расхождения дисперсий* (т. е. невыполнения гипотезы (8.4в)). Поэтому если мы хотим понять, за счет чего обнаружилась неоднородность анализируемых выборок, то необходимо произвести дополнительно проверку однородности дисперсий, т. е. гипотезы (8.4в). Эта же задача может являться и самостоятельной (автономной) целью исследования.

Критерий дисперсионного анализа является обобщением t -критерия на случай более чем двух сравниваемых совокупностей (т. е. $l > 2$). В этом случае критерий принадлежности нескольких выборок (8.3) из нормальных генеральных совокупностей с одинаковой (но неизвестной) дисперсией к одной общей совокупности основан на критической статистике

$$\gamma^{(n)} = \frac{\frac{1}{l-1} \sum_{j=1}^l n_j (\bar{x}_j(n_j) - \bar{x})^2}{\bar{s}^2}, \quad (8.26)$$

$\bar{x} = \sum_{j=1}^l n_j \bar{x}_j(n_j) / (n_1 + \dots + n_l)$ — общее выборочное среднее значение, а

$$\bar{s}^2 = \frac{1}{n_1 + \dots + n_l - l} \sum_{j=1}^l n_j s_j^2(n_j). \quad (8.27)$$

Показано, что если в рамках сформулированных выше условий справедлива гипотеза однородности (8.46) (а это будет означать, в силу нормальности сравниваемых совокупностей и одинаковости их дисперсий, и справедливость более сильной гипотезы (8.4а)), то статистика (8.26) подчиняется F -распределению с числом степеней свободы числителя, равным $l - 1$, и числом степеней свободы знаменателя, равным $n_1 + n_2 + \dots + n_l - l$. Поэтому, если окажется, что вычисленная по формуле (8.26) статистика

$$\gamma^{(n)} > F_{\alpha}(l - 1; \sum_{j=1}^l n_j - l), \quad (8.28)$$

то гипотезу (8.4а) об однородности выборок (8.3) следует отвергнуть (в правой части (8.28) $F_{\alpha}(l-1; \sum_{j=1}^l n_j - l) - 100\alpha\%$ -ная точка F -распределения с указанными выше числами степеней свободы числителя и знаменателя).

F -критерий однородности дисперсий. Этот критерий предназначен для проверки гипотезы однородности дисперсий (8.4в) в *двух нормальных* генеральных совокупностях. Он основан на использовании критической статистики

$$\gamma^{(n)} = \frac{\frac{1}{n_1-1} n_1 s_1^2(n_1)}{\frac{1}{n_2-1} n_2 s_2^2(n_2)}, \quad (8.29)$$

где все обозначения соответствуют ранее введенным.

При анализе статистики нормального закона (см. п. 6.2.8, Следствие 5, формула (6.38)) нами было показано, что в условиях справедливости гипотезы (8.4в) статистика (8.29) должна подчиняться F -распределению с числами степеней свободы числителя и знаменателя, равными соответственно $n_1 - 1$ и $n_2 - 1$. Поэтому при заданном уровне значимости критерия α определяем $100(1 - \alpha/2)\%$ -ную и $100\alpha/2\%$ -ную точки $F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$ и $F_{\alpha/2}(n_1 - 1, n_2 - 1)$ (см. таблицы в приложении 1). Если окажется, что

$$F_{1-\alpha/2}(n_1 - 1, n_2 - 1) \leq \gamma^{(n)} \leq F_{\alpha/2}(n_1 - 1, n_2 - 1),$$

то гипотеза однородности дисперсий не отвергается (и отвергается при всех других значениях статистики (8.29)).

Пример 8.3. Реклама утверждала, что из двух типов пластиковых карт «American Exspress» и «Visa» богатые люди предпочитают первый. Другими словами, среднемесячные платежи одного среднестатистического обладателя «American Exspress» существенно (*статистически значимо*) превышают среднемесячные платежи одного среднестатистического обладателя карты «Visa». С целью статистической проверки этого утверждения были обследованы среднемесячные платежи 32 обладателей «American» ($n_1 = 32$) и 30 обладателей «Visa» ($n_2 = 30$). В результате первичной статистической обработки этих двух выборок были получены следующие значения выборочных характеристик:

$$\begin{array}{l|l} \bar{x}_1(n_1) = \$563 & \bar{x}_2(n_2) = \$485 \\ s_1^2(n_1) = 31684 & s_2^2(n_2) = 38416. \end{array}$$

Требуется проверить гипотезу (8.4б) с уровнем значимости $\alpha = 0,05$.

Предварительный анализ законов распределения месячных расходов как среди обладателей «American», так и среди обладателей «Visa» показал, что и тот и другой з.р.в. достаточно хорошо описываются *нормальной моделью*. Но перед тем, как использовать в решении поставленной задачи t -критерий (критерий Стьюдента), необходимо убедиться в однородности дисперсий в анализируемых совокупностях, т.е. проверить гипотезу (8.4в) при $l = 2$. С этой целью воспользуемся *F-критерием*. Для экономии места таблицы процентных точек составлены только для случая, когда числитель больше знаменателя, поэтому, если в статистике (8.31) числитель оказывается меньше знаменателя, то следует изменить нумерацию выборок. Поэтому в нашем случае выборкой № 1 будем считать выборку из совокупности пользователей «Visa» (т.е. $n_1 = 30$).

Итак, в нашем примере отношение (8.29) оказывается равным 1,22. Из таблиц процентных точек находим пятипроцентную точку $F_{0,05}(29; 31) \approx 1,84$, а поскольку $1,22 < 1,84$, то мы *имеем основание принять допущение о равенстве дисперсий в анализируемых совокупностях в качестве статистически проверенной рабочей гипотезы*.

Теперь мы имеем основания воспользоваться критерием Стьюдента для проверки гипотезы о равенстве средних значений.

Вычисление критической статистики $\gamma^{(n)}$ по формуле (8.24) дает:

$$\gamma^{(n)} = \frac{563 - 485}{\sqrt{\frac{1}{30+32-2}(30 \cdot 196^2 + 32 \cdot 178^2)} \left(\frac{1}{30} + \frac{1}{32}\right)} = 1,61.$$

Из таблиц процентных точек находим $t_{0,025}(30 + 32 - 2) = 2,00$. И поскольку $|\gamma^{(n)}| = 1,61 < 2,00$, то мы делаем вывод о непротиворечивости гипотезы однородности средних значений имеющимся в нашем распоряжении статистическим данным. Следовательно, утверждения рекламы о более высокой состоятельности обладателей «American Express» не имели объективных оснований.

Критерий Бартлетта позволяет проверять гипотезу об однородности дисперсий в нескольких (в том числе *более чем в двух*) нормальных генеральных совокупностях. Критическая статистика критерия предложена Бартлеттом и имеет вид

$$\gamma^{(n)} = q \sum_{j=1}^l (n_j - 1) \ln \left(\frac{\bar{s}^2}{\bar{s}_j^2(n_j)} \right). \quad (8.30)$$

В формуле (8.30) величина $\bar{s}^2$ определяется соотношением (8.27),

$$\bar{s}_j^2(n_j) = n_j s_j^2(n_j) / (n_j - 1), \text{ а}$$

$$q = \left[1 + \frac{1}{3(l-1)} \left(\sum_{j=1}^l \frac{1}{n_j - 1} - \frac{1}{n_1 + \dots + n_l - l} \right) \right]^{-1}. \quad (8.31)$$

При $\min(n_1, \dots, n_l) > 3$ в условиях справедливости проверяемой гипотезы статистика (8.30) распределена приблизительно по закону χ^2 с $l - 1$ степенями свободы. Поэтому (при уровне значимости критерия α) если оказалось, что вычисленная по формуле (8.30) величина $\gamma_n > \chi_\alpha^2(l - 1)$, то гипотезу об однородности дисперсий в анализируемых нормальных генеральных совокупностях следует отвергнуть ($\chi_\alpha^2(l - 1)$ — 100%-ная точка «хи-квадрат» -распределения с $l - 1$ степенями свободы).

8.6.3. Проверка гипотез о числовых значениях параметров

Остановимся теперь на примерах статистических критериев, позволяющих проверять гипотезы вида (8.6). При этом важные статистические критерии такого типа, относящиеся к *корреляционному анализу, моделям регрессии, анализу временных рядов, системам одновременных уравнений*, будут обсуждаться в контексте изложения этих тем соответственно в гл. 10, 11 и в гл. 2, 3 и 4 второго тома. В данном пункте мы опишем несколько примеров статистических критериев, предназначенных для проверки *простых основных гипотез* (чаще всего — при простых же альтернативах) относительно числовых значений параметров анализируемых законов распределения вероятностей, т. е. речь идет о проверке основной («нулевой») гипотезы

$$\begin{cases} H_0: \Theta = \Theta_0 \\ \text{против альтернативы} \\ H_1: \Theta = \Theta_1 \quad (\text{или } \Theta \neq \Theta_0), \end{cases} \quad (8.32)$$

где Θ_0 и Θ_1 — заданные числовые значения параметра, который участвует в модельном описании функции распределения вероятностей анализируемой случайной величины ξ (т. е. $P\{\xi < x\} = F_\xi(x; \Theta)$).

Критерии проверки гипотез о числовом значении параметра p биномиального распределения. Рассматриваемая задача относится к анализу результатов серии n независимых испытаний Бернулли (см. п. 3.1.1). При этом в имеющейся у нас выборке объема n интересующее нас событие произошло x раз. Можно интерпретировать эту серию (выборку) как единственное наблюдение (n, p) -биномиальной случайной величины в

ситуации, когда параметр n известен, а параметр p нет. Далее речь пойдет о проверке гипотез типа (8.32), где роль параметра Θ играет параметр p , т.е. о проверке (по имеющимся у нас результатам наблюдения) гипотез о численном значении параметра p биномиального распределения при различных альтернативах.

Принимая во внимание утверждение леммы Неймана–Пирсона о том, что *критерии отношения правдоподобия* являются наиболее мощными среди всех других возможных критериев (см. п. 8.3.1), попробуем вывести критическую статистику критерия, руководствуясь именно этим принципом. При этом нам удобнее будет, как это часто бывает при работе с функциями правдоподобия, не само отношение правдоподобия, а его логарифм (монотонность этого преобразования обеспечивает нужный результат).

Функция правдоподобия биномиального закона с параметрами n и p при единственном наблюдении x , как известно (см. п. 3.1.1), имеет вид:

$$L(x | p; n) = C_n^x p^x (1 - p)^{n-x}.$$

Так что критическая статистика критерия $\gamma^{(n)}$, определяемая логарифмом отношения правдоподобия при *произвольном* значении параметра p по отношению к *основному гипотетическому* p_0 , будет иметь вид:

$$\gamma^{(n)} = \ln \frac{L(x | p; n)}{L(x | p_0; n)} = x \ln \left(\frac{p}{1-p} \cdot \frac{1-p_0}{p_0} \right) + n \ln \frac{1-p}{1-p_0}. \quad (8.33)$$

Из смысла статистики γ_n ясно, что «достаточно большие» ее значения сигнализируют о большей правдоподобности *конкурирующей* гипотезы « $H: p \neq p_0$ », т.е. о необходимости отвергнуть основную гипотезу « $H_0: p = p_0$ ».

Для того чтобы построить критерий при заданном значении уровня значимости α , нужно уметь назначить такое пороговое значение c_α , при котором

$$P\{\gamma^{(n)} \geq c_\alpha | H_0\} = \alpha. \quad (8.34)$$

А для того, чтобы вычислить ошибку второго рода β или мощность критерия $1 - \beta$, нужно уметь вычислять вероятность

$$1 - \beta = P\{\gamma^{(n)} \geq c_\alpha | H_1\}. \quad (8.35)$$

Из (8.33) следует, что обе эти задачи решаются, если мы будем знать распределение случайной величины x как при условии справедливости «нулевой» гипотезы H_0 (т.е. при значении параметра, равном заданной величине p_0), так и при условии справедливости любой альтернативы

(т.е. при любом другом значении p анализируемого параметра). Но признак x , по построению, есть (n, p) -биномиально распределенная случайная величина со значением параметра p , определяемым в зависимости от того, в условиях справедливости какой из гипотез мы ее рассматриваем.

Поэтому в дальнейшем мы в качестве критической статистики $\tilde{\gamma}^{(n)}$ будем рассматривать (n, p) -биномиально распределенную случайную величину или любую удобную для нас ее линейную комбинацию. Правда, при переходе от непосредственно вычисленной критической статистики $\gamma^{(n)}$, задаваемой соотношением (8.33), к ее линейным комбинациям смысл неравенств в фигурных скобках (8.34) и (8.35) может меняться на противоположный. Рассмотрим подробнее три возможных варианта анализируемой задачи.

Вариант 1. Проверяется простая гипотеза $H_0: p = p_0$ при простой альтернативе $H_1: p = p_1$, причем $p_1 > p_0$. Тогда смысл неравенств в (8.34) и (8.35) сохраняется и при замене $\gamma^{(n)}$ на x , а именно по заданному уровню значимости α требуется найти такое c_α , что

$$P\{x \geq c_\alpha \mid p_0\} = \alpha. \quad (8.34')$$

Но

$$P\{x \geq c_\alpha \mid p_0\} = \sum_{k=c_\alpha}^n C_n^k p_0^k (1-p_0)^{n-k} = \alpha. \quad (8.34'')$$

Следовательно, требуется решить уравнение (8.34'') относительно c_α . Обычно для этого пользуются либо нормальным, либо пуассоновским приближением, а именно:

• если гипотетическая величина p_0 «не слишком близка» к нулю или единице (скажем, $0,10 \leq p_0 \leq 0,90$), а число наблюдений n составляет хотя бы несколько десятков, то используют теорему Муавра–Лапласа (см. п. 4.3.1) о приближенной (асимптотической) стандартной нормальности случайной величины

$$\tilde{\gamma}^{(n)} = \frac{\frac{x}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}.$$

Тогда при $p = p_0$

$$P\{x \geq c_\alpha \mid p_0\} = P\left\{\frac{\frac{x}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq \frac{\frac{c_\alpha}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \mid p_0\right\} \quad (8.34''')$$

$$\approx 1 - \Phi\left(\frac{\frac{c_\alpha}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right) = \alpha.$$

Следовательно, аргумент функции стандартного нормального распределения является ничем иным, как квантилем уровня $1 - \alpha$ этого распределения ($u_{1-\alpha}$). Так что определив из таблиц величину $u_{1-\alpha}$, имеем

$$\left(\frac{c_\alpha}{n} - p_0\right) / \sqrt{\frac{p_0(1-p_0)}{n}} = u_{1-\alpha}. \quad (8.36)$$

Из (8.36) определяем величину c_α , на которой основано правило проверки гипотезы H_0 : если окажется, что $x \geq c_\alpha$, то гипотеза H_0 отвергается (с вероятностью ошибки, приблизительно равной α).

Ошибка второго рода этого критерия подсчитывается также с использованием нормальной аппроксимации биномиального закона, но при значении параметра $p = p_1$, а именно:

$$\beta = P\{x < c_\alpha \mid H_1\} = P\{x < c_\alpha \mid p_1\}$$

$$= \sum_{k=0}^{c_\alpha-1} C_n^k p_1^k (1-p_1)^{n-k} \approx \Phi\left(\left(\frac{c_\alpha}{n} - p_1\right) / \sqrt{\frac{p_1(1-p_1)}{n}}\right); \quad (8.37)$$

• если гипотетическая величина p_0 близка к 0 или единице (т.е. $p_0 < 0,10$ или $p_0 > 0,9$), а число наблюдений, как и в предыдущем случае, составляет хотя бы несколько десятков, то для вычисления вероятностей событий вида $\{x = k\}$, где x — (n, p) -биномиально распределенная случайная величина, лучше пользоваться пуассоновской аппроксимацией, т.е.

$$P\{x = k\} \approx \frac{(np)^k}{k!} e^{-np}. \quad (8.38)$$

Использование таблиц пуассоновского распределения с параметром $\lambda = np$ позволит по схеме, аналогичной предыдущему случаю, определить сначала c_α из условия

$$P\{x \geq c_\alpha \mid p_0\} \approx \sum_{k=c_\alpha}^{\infty} \frac{(np_0)^k}{k!} e^{-np_0} = 1 - F(c_\alpha \mid \lambda = np_0) = \alpha, \quad (8.39)$$

а затем вычислить вероятность ошибки второго рода

$$\beta = P\{x < c_\alpha \mid p_1\} = F(c_\alpha \mid \lambda = np_1) \quad (8.40)$$

(в соотношениях (8.39) и (8.40) функция $F(x \mid \lambda)$ обозначает функцию распределения пуассоновского закона с параметром λ).

Вариант 2. Проверяется простая гипотеза $H_0: p = p_0$ при простой альтернативе $H_1: p = p_1$, причем $p_1 < p_0$. Схема решения в точности повторяет предыдущую с заменой смысла неравенства в соотношении (8.34') на противоположный. А именно критическая константа c_α по заданному уровню значимости α определяется из условия

$$P\{x \leq c_\alpha \mid p_0\} = \alpha. \quad (8.36'a)$$

Соответственно меняется на противоположный и смысл неравенств в фигурных скобках соотношений (8.34''), (8.34'''), (8.37), (8.39) и (8.40) с соответствующими изменениями в последующих выкладках.

Вариант 3. Проверяется простая гипотеза $H_0: p = p_0$ против сложной альтернативы $H_1: p \neq p_0$. В этом случае одинаково «неестественными» (с точки зрения справедливости гипотезы H_0) будут большие отклонения x от np_0 как в одну, так и в другую сторону. Это отразится на схеме построения критерия следующим образом:

- при использовании нормальной аппроксимации следует находить константу $c_{\alpha/2}$, заменив в правой части (8.34''') величину α на $\alpha/2$, тогда решение об отклонении гипотезы H_0 будет приниматься в случае

$$\left| \frac{x}{n} - p_0 \right| / \sqrt{\frac{p_0(1-p_0)}{n}} > u_{1-\alpha/2}, \quad (8.41)$$

где u_q , как и прежде, квантиль уровня q стандартного нормального распределения;

- при использовании пуассоновской аппроксимации потребуются вычислить две критические константы: $100(1-\alpha/2)\%$ -ную и $100\alpha/2\%$ -ную точки (соответственно $c_{1-\alpha/2}$ и $c_{\alpha/2}$) пуассоновского распределения с параметром $\lambda = np_0$; решение об отклонении гипотезы. Но будет приниматься в случаях $x < c_{1-\alpha/2}$ и $x > c_{\alpha/2}$.

Пример 8.4. Из массового стационарно функционирующего производства извлекается контрольная выборка изделий объемом 100 единиц ($n = 100$) с целью проверки гипотезы H_0 о том, что доля брака производства составляет $p_0 = 0,05$, против гипотезы H_1 о том, что эта доля равна $p_1 = 0,10$. Построить наиболее мощный критерий проверки гипотезы H_0 с уровнем значимости α , не превышающим 0,05, и вычислить его мощность.

Решение. Условие задачи соответствует вышеописанному варианту 1, а соотношение величин p_0 и n говорит за то, что следует воспользоваться пуассоновской аппроксимацией при определении критической константы $c_{0,05}$. В соответствии с (8.34'') и (8.39) с помощью таблиц пуассоновского распределения с параметром $\lambda = np_0 = 100 \cdot 0,05 = 5$ определяем:

$$P\{x \leq 8\} \approx F(9 | \lambda = 5) = 0,932;$$

$$P\{x \leq 9\} \approx F(10 | \lambda = 5) = 0,968.$$

Так что выбрав $c_\alpha = 10$, имеем величину уровня значимости критерия, равную $\alpha = 1 - F(10 | \lambda = 5) = 0,032 < 0,05$. Таким образом, критерий задается критической статистикой — x -числом дефектных изделий, обнаруженных в выборке объема $n = 100$, и критической константой $c_\alpha = 9$: если в выборке объема $n = 100$ окажется более 9 дефектных изделий, то гипотезу о доле брака производства, равной 0,10, следует предпочесть гипотезе о доле брака производства, равной 0,05.

Подсчет мощности критерия (с учетом формулы (8.40)) дает:

$$1 - \beta = 1 - P\{x < c_\alpha | p = 0,10\} \approx 1 - \sum_{k=0}^9 \frac{10^k}{k!} e^{-10} = 0,542.$$

Таким образом, ошибка второго рода, равная $\beta = 0,458$, оказывается весьма большой.

Пример 8.5. Независимому статистику поручено проверить информацию маркетинговой службы некоторого туристического бюро о том, что 70% клиентов выбирают в качестве формы обслуживания полупансион. Статистик провел опрос 150 случайно выбранных туристов, из них полупансион предпочли 84 человека. К какому выводу пришел статистик при проверке гипотезы $H_0: p = 0,70$ при альтернативе $H_1: p \neq 0,70$ при уровне значимости критерия $\alpha = 0,05$?

Решение. Условия задачи соответствуют вышеописанному варианту 3, а соотношение величин $p_0 = 0,70$ и $n = 150$ говорит за то, что можно воспользоваться нормальной аппроксимацией. Расчет величин, участвующих в соотношении (8.41), приводит к следующим результатам:

$$\left(\frac{x}{n} - p_0\right) / \sqrt{\frac{p_0(1-p_0)}{n}} = \left(\frac{84}{150} - 0,7\right) / \sqrt{\frac{0,7 \cdot 0,3}{150}} = -3,74.$$

А поскольку $u_{1-\alpha/2} = u_{0,975} = 1,96$, то неравенство (8.41) имеет место и, следовательно, есть основания подвергнуть сомнению данные маркетинговой службы, т.е. отвергнуть гипотезу H_0 .

Критерии проверки гипотез о среднем значении нормальной генеральной совокупности. Все критические статистики, которые используются ниже для построения критериев проверки гипотез о среднем значении нормально распределенной случайной величины, получаются с помощью реализации *принципа отношения правдоподобия* (8.7), описанного в п. 8.3.1. А потому и основанные на них критерии, в соответствии с леммой Неймана–Пирсона, являются в определенном смысле наилучшими (наиболее мощными). Мы не будем здесь подробно расписывать логарифмы отношения правдоподобия для конкурирующих гипотетических нормальных законов: думаем, у читателя, овладевшего предыдущим материалом учебника, не возникнет принципиальных трудностей при самостоятельном выводе критических статистик в данном случае. Поэтому ограничимся лишь сводкой результатов.

Итак, мы имеем случайную выборку $(x_1, x_2, \dots, x_n)$ из (a, σ^2) -нормальной генеральной совокупности. Рассмотрим различные варианты постановок задач по статистической проверке гипотез о числовом значении параметра $a = E\xi$.

1) $H_0: a = a_0$ при альтернативе $H_1: a > a_0$; σ^2 известна. Критическая статистика $\gamma^{(n)} = (\bar{x} - a_0)\sqrt{n}/\sigma$.

Правило принятия решения: если $\gamma^{(n)} > u_{1-\alpha}$, то гипотеза H_0 отвергается с вероятностью ошибки, равной α (здесь и далее u_q -квантиль уровня q стандартного нормального распределения).

Обоснование: статистика $\gamma^{(n)}$ в соответствии с теоремой Фишера (см. п. 6.2) подчиняется стандартному нормальному закону в условиях справедливости гипотезы H_0 .

2) $H_0: a = a_0$ при альтернативе $H_1: a < a_0$; σ^2 известна.

Критическая статистика: $\gamma^{(n)} = (\bar{x} - a_0)\sqrt{n}/\sigma$.

Правило принятия решения: если $\gamma^{(n)} < -u_{1-\alpha}$, то гипотеза H_0 отвергается с вероятностью ошибки, равной α .

Обоснование: то же, что в предыдущем пункте.

3) $H_0: a = a_0$ при альтернативе $H_1: a \neq a_0$; σ^2 известна.

Критическая статистика: $\gamma^{(n)} = (\bar{x} - a_0)\sqrt{n}/\sigma$.

Правило принятия решения: если $|\gamma^{(n)}| > u_{1-\alpha/2}$, то гипотеза H_0 отвергается с вероятностью ошибки, равной α .

Обоснование: то же, что в предыдущих пунктах.

4) $H_0: a = a_0$ при альтернативах $H_1: a > a_0$; или $H_1: a < a_0$; или $H_1: a \neq a_0$; σ^2 неизвестна. Схемы критериев для каждой из альтернатив повторяют соответствующие схемы для известной дисперсии σ^2 со следующими изменениями, касающимися критической статистики и правила

принятия решения.

Критическая статистика: $\gamma^{(n)} = (\bar{x} - a_0)\sqrt{n-1}/s$.

Правило принятия решения: гипотеза H_0 отвергается, если

$$\gamma^{(n)} > t_{\alpha}(n-1) \quad \text{при альтернативе} \quad H_1: a > a_0;$$

$$\gamma^{(n)} < -t_{\alpha}(n-1) \quad \text{при альтернативе} \quad H_1: a < a_0;$$

$$|\gamma^{(n)}| > t_{\alpha/2}(n-1) \quad \text{при альтернативе} \quad H_1: a \neq a_0$$

(здесь $t_q(n-1)$ — 100q%-ная точка распределения Стьюдента с $n-1$ степенью свободы).

Обоснование: в соответствии со Следствием 2 теоремы Фишера (см. п. 6.2, (6.30)) статистика $\gamma^{(n)}$ распределена по закону Стьюдента с $n-1$ степенью свободы.

Критерий проверки гипотезы о значении дисперсии нормальной генеральной совокупности. По случайной выборке $(x_1, x_2, \dots, x_n)$ из (a, σ^2) -нормальной генеральной совокупности получена выборочная дисперсия $s^2(n)$. Требуется проверить гипотезу (при уровне значимости критерия α) $H_0: \sigma^2 = \sigma_0^2$, где σ_0^2 — некоторое конкретное числовое значение. При проверке этой гипотезы используют критическую статистику

$$\gamma^{(n)} = \frac{ns^2(n)}{\sigma_0^2}, \quad (8.42)$$

которая в соответствии с теоремой Фишера (см. (6.27) в п. 6.2.8) в условиях справедливости гипотезы H_0 распределена по закону χ^2 с $n-1$ степенью свободы.

Далее, как и в случаях проверки гипотез о параметре биномиального закона p или о значении среднего нормальной генеральной совокупности a , в зависимости от конкурирующей гипотезы H_1 выбирают правостороннюю, левостороннюю или двустороннюю критическую область:

1) если $H_1: \sigma^2 = \sigma_1^2 > \sigma_0^2$, то гипотезу H_0 отвергают (с вероятностью ошибки α) в случае $ns^2/\sigma_0^2 > \chi_{\alpha}^2(n-1)$;

2) если $H_1: \sigma^2 = \sigma_1^2 < \sigma_0^2$, то гипотезу H_0 отвергают (с вероятностью ошибки α) в случае $ns^2/\sigma_0^2 < \chi_{1-\alpha}^2(n-1)$;

3) если $H_1: \sigma^2 \neq \sigma_0^2$, то гипотеза H_0 отвергается (с вероятностью ошибки α) при $ns^2/\sigma_0^2 < \chi_{1-\alpha/2}^2(n-1)$, а также при $ns^2/\sigma_0^2 > \chi_{\alpha/2}^2(n-1)$. Здесь, как обычно, $\chi_q^2(m)$ обозначает 100q%-ную точку χ^2 -распределения с m степенями свободы.

ВЫВОДЫ

1. Процедура обоснованного сопоставления высказанного исследователем предположительного утверждения (гипотезы) относительно природы или величины неизвестных параметров рассматриваемой стохастической системы с имеющимися в его распоряжении результатами наблюдения, сопровождаемая количественной оценкой степени достоверности получаемого вывода, осуществляется с помощью того или иного статистического критерия и называется статистической проверкой гипотез.

2. По своему прикладному содержанию гипотезы, высказываемые в ходе статистического анализа и моделирования, подразделяют на следующие типы:

- об общем виде закона распределения исследуемой случайной величины;

- об однородности двух или нескольких обрабатываемых выборок;

- о числовых значениях параметров исследуемой генеральной совокупности;

- об общем виде зависимости, существующей между компонентами исследуемого многомерного признака;

- о независимости и стационарности ряда наблюдений.

3. Все статистические критерии строятся по общей логической схеме. Построить статистический критерий — это значит: а) определить тип проверяемой гипотезы; б) предложить и обосновать конкретный вид функции от результатов наблюдения (критической статистики $\gamma^{(n)}$), на основании значений которой принимается окончательное решение; в) указать такой способ выделения из области возможных значений критической статистики γ^n области Γ_n^H отклонения проверяемой гипотезы H_0 , чтобы было соблюдено требование к величине вероятности ошибочного отклонения гипотезы H_0 (т.е. к уровню значимости критерия α).

4. «Качество» статистического критерия характеризуется уровнем значимости α , мощностью $1 - \beta$, свойствами несмещенности и состоятельности. В состоятельных критериях можно добиваться сколько угодно малых величин ошибок первого и второго рода (α и β) лишь за счет увеличения объема выборки n , на основании которой принимается решение. При фиксированном объеме выборки можно делать сколь угодно малой лишь одну из ошибок (α или β), что сопряжено с неизбежным увеличением другой.

5. Наряду с классической схемой наблюдения, когда объем выборки n заранее зафиксирован, в практике статистических исследований используется и *последовательная* схема наблюдения, при которой на каждом из последовательно во времени проводимых этапов наблюдения принимает

ся одно из трех решений: «принять гипотезу H_0 », «отклонить гипотезу H_0 », «не принимать окончательного решения и продолжить наблюдения». При этом выбор решения ставится в зависимость от результатов всех предыдущих наблюдений, а число наблюдений ν , произведенных до момента принятия окончательного решения, оказывается величиной случайной.

6. Оптимальные последовательные критерии отношения правдоподобия (критерий Вальда, обобщенный последовательный критерий и др.) оказываются более экономными по затратам на наблюдения, на основании которых можно различить проверяемые гипотезы с заданной точностью (α, β) . Исследования показали, что, применяя последовательные критерии, можно добиваться двух-, трех- и даже четырехкратного снижения необходимого числа наблюдений по сравнению с классическими оптимальными критериями.

Раздел III

Методы Прикладной Статистики

- Глава 9. Введение в прикладной статистический анализ
- Глава 10. Статистическое исследование зависимостей (основные понятия и постановки задач)
- Глава 11. Корреляционный анализ многомерной генеральной совокупности
- Глава 12. Распознавание образов и типологизация объектов в социально-экономических исследованиях (методы классификации)
- Глава 13. Снижение размерности исследуемого многомерного признака и отбор наиболее информативных показателей

ГЛАВА 9. ВВЕДЕНИЕ В ПРИКЛАДНОЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ

9.1. Назначение и содержание прикладной статистики

Во «Введении» к учебнику (см. п. В.1.2) мы кратко обсудили вопросы, вынесенные в заголовок этого пункта, дали определение прикладной статистики как самостоятельной научной дисциплины, попытались понять «взаимоотношения» прикладной статистики с теорией вероятностей и математической статистикой, с одной стороны, и с эконометрикой — с другой. В частности, было установлено, что теория вероятностей и математическая статистика являются по отношению к прикладной статистике и эконометрике разработчиками и поставщиками существенной части используемого в них математического аппарата. Однако *доводка и развитие этого аппарата, подчиненные требованиям и специфике различного рода приложений, производятся уже в рамках дисциплин «прикладная статистика» и «эконометрика»*. Причем эконометрика (наряду с другими сферами человеческой деятельности, в которых используются математико-статистические методы) является активным поставщиком специального класса постановок задач для прикладной статистики. Далее развиваются и комментируются некоторые из тезисно намеченных во «Введении» к учебнику положений, относящихся к назначению и содержанию прикладной статистики.

9.1.1. Два подхода к интерпретации и анализу исходных статистических данных

Первый подход (вероятностно-статистический) развивается в рамках классической математической статистики (т.е. в условиях хотя бы приблизительного выполнения требований статистического ансамбля,

см. п. В.2.1) и предусматривает возможность вероятностной интерпретации анализируемых данных и получаемых в результате этого анализа статистических выводов. При подобной (вероятностной) интерпретации исходных статистических данных в поле зрения исследователя одновременно попадают две совокупности объектов: *реально наблюдаемая*, статистически представленная рядом наблюдений (т.е. *выборка*), и теоретически домысливаемая (так называемая *генеральная совокупность*). Основные свойства и характеристики выборки, называемые эмпирическими (или выборочными), могут быть проанализированы и вычислены по имеющимся статистическим данным. Основные свойства и характеристики генеральной совокупности, называемые теоретическими, не известны исследователю, но назначение математико-статистических методов как раз в том и состоит, чтобы с их помощью получить как можно более точное представление об этих теоретических свойствах и характеристиках по соответствующим свойствам и характеристикам выборок (см. гл. 6).

Важнейшим моментом в успехе статистического анализа и моделирования таких данных является удачное решение **проблемы их генезиса**, т.е. *правильный выбор вероятностной модели механизма генерации этих данных* (набор наиболее распространенных моделей подобного рода описан в гл. 3). Именно отправляясь от решения проблемы генезиса анализируемых статистических данных, мы можем обоснованно ответить на вопросы:

- как наилучшим образом выбрать метод статистической обработки этих данных, например, построить наиболее точную оценку неизвестного среднего значения анализируемой генеральной совокупности?
- каким должен быть общий вид модели, описывающий ту или иную зависимость между анализируемыми признаками?

В обоих случаях **критерий качества** (оценки, степени адекватности модели) определится на основе *принципа максимального правдоподобия* имеющихся у нас наблюдений, который в свою очередь базируется на знании модели з.р.в. этих наблюдений (см. пп. 7.5.1, 8.3.1).

В принципиально иной ситуации оказывается исследователь, если он не располагает никакими априорными сведениями о вероятностной природе анализируемых данных, или если эти данные вообще не могут быть интерпретированы как *выборка из генеральной совокупности*. Тогда при выборе критерия качества (метода оценивания или степени адекватности конструируемой модели) исследователь вынужден опираться на соображения конкретно-содержательного плана: как именно получены анализируемые данные и какова конечная прикладная цель их анализа. Поскольку эти соображения основаны на обычной логике и реализуются, как

правило, в виде критерия некоторого *алгебраического* вида, то и соответствующий подход принято называть **логико-алгебраическим**. Очевидно, в рамках логико-алгебраического подхода (в отличие от вероятностно-статистического) исследователь не может претендовать на:

- интерпретацию исходных статистических данных в качестве выборки из некоторой (теоретически домысливаемой) генеральной совокупности;
- использование вероятностных моделей для построения и выбора наилучших методов статистической обработки или наилучшего вида конструируемой модели;
- вероятностную интерпретацию выводов, основанную на статистическом анализе исходных данных.

В этом заключается главное различие двух возможных подходов к статистическому анализу исходных данных. Однако и в том и в другом подходе выбор наилучшего из всех возможных методов анализа и моделирования данных производится в соответствии с некоторым критерием *качества метода* или степени адекватности модели. Различие описываемых подходов проявляется здесь в *способе обоснования* выбора этого критерия качества, а также в интерпретации самого критерия и получаемых статистических выводов. Но после того, как выбор конкретного вида оптимизируемого критерия качества осуществлен, математические средства решения задачи статистического анализа и моделирования данных оказываются общими для обоих подходов: и в том и в другом случае с целью оптимизации выбранного критерия качества исследователь использует методы решения экстремальных задач. Правда, на заключительном этапе — на этапе осмысления и интерпретации полученных статистических выводов — каждый из подходов снова имеет свою специфику.

Таким образом, общим для обоих описываемых подходов является наличие исходной статистической информации на «входе» задачи и необходимость наилучшего (в смысле оптимизации некоторого критерия качества) способа статистического анализа или моделирования этой информации с целью получения научных или практических выводов «на выходе».

Именно эта общая логическая схема и положена в основу методологического принципа разработки инструментария прикладной статистики. Это позволило синтезировать два описанных выше подхода к статистическому анализу и моделированию. В результате прикладная статистика объединила в себе как *прикладные вероятностно-статистические методы* многомерного статистического анализа, включая модели регрессии (существенно опираясь при этом на результаты математической статистики, касающиеся методов статистического оценивания и статистиче-

ской проверки гипотез, см. гл. 7 и 8), так и *логико-алгебраические методы анализа данных*, понимаемого в узком смысле, т. е. исключающего использование в своих построениях вероятностных рассуждений и моделей (достаточно полное представление об этих методах можно получить, например, по книге [Диде Э. и др.]).

9.1.2. Три центральные проблемы прикладной статистики

Перед тем как сформулировать три центральные проблемы прикладной статистики, следует остановиться на двух основных формах записи исходных статистических данных (и.с.д.). Первую, наиболее распространенную, форму представления и.с.д. обычно называют **матрицей (или таблицей) «объект-свойство»**. В своей наиболее общей записи эта матрица имеет вид

$$(и.с.д.)_1 = \begin{pmatrix} x_1^{(1)}(t) & x_1^{(2)}(t) & \dots & x_1^{(p)}(t) \\ x_2^{(1)}(t) & x_2^{(2)}(t) & \dots & x_2^{(p)}(t) \\ \dots & \dots & \dots & \dots \\ x_n^{(1)}(t) & x_n^{(2)}(t) & \dots & x_n^{(p)}(t) \end{pmatrix}, \quad (9.1)$$

$$t = t_1, t_2, \dots, t_N,$$

где $x_i^{(j)}(t_k)$ — значение j -го анализируемого признака, характеризующего состояние i -го объекта в момент времени t_k . Данные (9.1) образуют так называемую **пространственно-временную выборку**, при формировании которой статистическому обследованию подвергаются n объектов (как-то размещенных в *пространстве*), причем на каждом из объектов регистрируются значения p характеризующих его признаков в N последовательные моменты времени $t_1, t_2, \dots, t_N$. Очевидно, что запись (9.1) в действительности определяет **целую последовательность** (а именно N штук) матриц «объект-свойство». Можно также сказать, что данные вида (9.1) содержат n **реализаций p -мерного временного ряда** $(x^{(1)}(t), x^{(2)}(t), \dots, x^{(p)}(t))$.

Если мы располагаем так называемыми **одномоментными наблюдениями**, то это соответствует случаю $N = 1$ в общей записи (9.1); при этом для сокращения обозначений индекс времени t в записи (9.1) мы можем опускать, а получающаяся выборка

$$(и.с.д.)_1 = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(p)} \\ \dots & \dots & \dots & \dots \\ x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(p)} \end{pmatrix} \quad (9.1')$$

называется *пространственной статической*.

Другой частный случай и.с.д. мы получаем, если в (9.1) положить $n = 1$ (т.е. обследуется во времени *единственный* объект). Тогда речь идет об анализе единственной траектории p -мерного временного ряда, а при дополнительном условии $p = 1$, — об анализе одного временного ряда.

Заметим, что для экономических приложений наиболее типична ситуация, когда моменты времени $t_1, t_2, \dots, t_N$, в которые производится регистрация значений анализируемых признаков, являются *равноотстоящими*, т.е. $t_2 - t_1 = t_3 - t_2 = \dots = t_N - t_{N-1} = \Delta t$. В этом случае время удобнее считать и обозначать в числе «тактов» Δt . Соответственно тогда вместо $t_1, t_2, \dots, t_N$ мы будем иметь $t = 1, 2, \dots, N$.

В ряде ситуаций и в первую очередь в ситуациях, когда исходные статистические данные получают с помощью специальных опросов, анкет, экспертных оценок, возможны случаи, когда элементом первичного наблюдения является не состояние i -го объекта в момент t , а характеристика $\gamma_{ij}(t)$ попарного сравнения двух объектов (или признаков) соответственно с номерами i и j , отнесенная к моменту времени t .

Характеристика γ_{ij} определяет результат сопоставления объектов O_i и O_j в смысле некоторого анализируемого отношения: γ_{ij} может выражать меру сходства или различия объектов O_i и O_j , меру их связи или взаимодействия в каком-либо процессе (например, поток продукции отрасли « i » в отрасль « j »), геометрическое расстояние между объектами, отношение предпочтения (тогда, например, полагают $\gamma_{ij} = 1$, если объект O_i не хуже объекта O_j , и $\gamma_{ij} = 0$ в противном случае), наконец, статистическую меру взаимной коррелированности (если речь идет о *признаках*, то в качестве γ_{ij} могут рассматриваться, например, коэффициенты корреляции или ковариации признаков $x^{(i)}$ и $x^{(j)}$). Из данного выше пояснения смысла характеристик γ_{ij} следует, что в *большинстве случаев* (но не всегда!) они оказываются *симметричными*, т.е. $\gamma_{ij} = \gamma_{ji}$ (пример с взаимными потоками продукции из одной отрасли в другую как раз относится к исключениям из этого правила).

Таким образом, в описываемой ситуации исследователь располагает в качестве массива исходных статистических данных временной последовательностью **матриц парных сравнений** размера $n \times n$ (если рассматриваются характеристики парных сравнений *объектов*) или $p \times p$ (если

расстояний (близостей) задана матрицей (γ_{ij}) , $i, j = 1, 2, \dots, n$. Аналогично интерпретируется исходная информация в задаче классификации признаков с той лишь разницей, что каждый из признаков $x^{(j)}$ задается соответствующим (j -м) столбцом матрицы (9.1') или, в случае задания и.с.д. в форме (9.2), — j -ми строками и столбцами матрицы (9.2).

В дальнейшем, если это специально не оговорено, не будем разделять изложение этой проблемы на «объекты» и «признаки», поскольку все постановки задач и основная методологическая схема исследования здесь общие.

В общей (нестрогой) постановке *проблема классификации объектов заключается в том, чтобы всю анализируемую совокупность объектов* $O = \{O_i\}$ ($i = 1, \dots, n$), статистически представленную в виде матриц (9.1') или (γ_{ij}) , *разбить на сравнительно небольшое число (заранее известное или нет) однородных, в определенном смысле, групп или классов.*

Для формализации этой проблемы удобно интерпретировать анализируемые объекты в качестве точек в соответствующем признаковом пространстве. Если исходные данные представлены в форме матрицы (9.1'), то эти точки являются непосредственным геометрическим изображением многомерных наблюдений $X_1, X_2, \dots, X_n$ в p -мерном пространстве $Ox^{(1)}, \dots, Ox^{(p)}$ (т.е. $X_i = (x_i^{(1)}, \dots, x_i^{(p)})^T$). Если же исходные данные представлены в форме матрицы парных сравнений γ_{ij} , то исследователю не известны непосредственно координаты этих точек, но зато задана структура парных расстояний (близостей) между объектами. Естественно предположить, что геометрическая близость двух или нескольких точек в этом пространстве означает близость «физических» состояний соответствующих объектов, их однородность. Тогда проблема классификации состоит в разбиении анализируемой совокупности точек-наблюдений на сравнительно небольшое число (заранее известное или нет) классов таким образом, чтобы объекты, принадлежащие одному классу, находились бы на сравнительно небольших расстояниях друг от друга. Полученные в результате разбиения классы часто называют *кластерами* (таксонами, образами)¹, а методы их нахождения соответственно кластер-анализом, численной таксономией, распознаванием образов.

Статистическим методам классификации посвящена гл. 12 учебника.

¹ *Cluster* (англ.) — гроздь, пучок, скопление, группа элементов, характеризуемых каким-либо общим свойством. *Taxon* (англ.) — систематизированная группа любой категории (термин биологического происхождения). Название «кластер-анализ» для совокупности методов решения задач такого типа было впервые введено, по-видимому, Трайном в 1939 г. (см. : Т р у о п R. C. Cluster Analysis. II Ann. Arb. Edw. Brothers. — 1939).

В зависимости от наличия и характера априорных сведений о природе искоемых классов и от конечных прикладных целей исследования приходится обращаться либо к методам *дискриминантного анализа* (они описаны в п. 12.3), либо к методам *расщепления смесей вероятностных распределений* или процедурам *кластер-анализа* (и те и другие описаны в п. 12.4).

Проблема III. Снижение размерности исследуемого признакового пространства с целью лаконичного объяснения природы анализируемых многомерных данных. Возможность лаконичного описания анализируемых многомерных данных основана на априорном допущении, в соответствии с которым существует небольшое (в сравнении с числом p исходных анализируемых признаков $x^{(1)}, x^{(2)}, \dots, x^{(p)}$) число p' *признаков-детерминант* (главных компонент, общих факторов, наиболее информативных объясняющих переменных), с помощью которых могут быть достаточно точно описаны как сами наблюдаемые переменные анализируемых объектов, т. е. все элементы матриц (9.1) и (9.2), так и *определяемые этими переменными свойства (характеристики) анализируемой совокупности*. При этом упомянутые признаки-детерминанты могут находиться среди исходных признаков, а могут быть *латентными*, т. е. непосредственно статистически не наблюдаемыми, но восстанавливаемыми по исходным данным вида (9.1), (9.1') или (9.2). Гениальный пример практической реализации этой идеи дает нам периодическая система элементов Менделеева: в этом случае роль идеально информативного *единственного признака-детерминанта* играет, как известно, заряд атомного ядра.

Необходимость снижения размерности исследуемого признакового пространства с целью лаконичного объяснения природы анализируемых многомерных данных может быть продиктована различными прикладными задачами статистического анализа и моделирования. Ниже кратко формулируются несколько типовых задач такого рода.

Отбор наиболее информативных показателей (включая выявление латентных факторов). Речь идет об отборе из исходного (априорного) множества признаков $X = (x^{(1)}, \dots, x^{(p)})^T$ или о построении в качестве некоторых комбинаций исходных признаков относительно небольшого числа p' переменных $Z(X) = (z^{(1)}(X), \dots, z^{(p')}(X))^T$, которые обладали бы свойством наибольшей информативности в смысле, определенном, как правило, некоторым специально подобранным для каждого конкретного типа задач *критерием информативности* $I_{p'}(Z)$. Так, например, если критерий $I_{p'}(Z)$ «настроен» на достижение максимальной точности регрессионного прогноза некоторого результирующего количественного показателя y по известным значениям предикторных перемен-

ных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, то речь идет о наилучшем подборе наиболее существенных предикторов в модели регрессии (см. п. 2.6 тома 2). Если же критерий $I_{p'}(Z)$ устроен таким образом, что его оптимизация обеспечивает наивысшую точность решения задачи отнесения объекта к одному из классов по значениям X его описательных признаков, то речь идет о построении системы типобразующих признаков в задачах классификации (см. п. 12.3 и 12.4) или о выявлении и интерпретации некоторой сводной (латентной) характеристики изучаемого свойства (см. п. 13.5). Наконец, критерий $I_{p'}(Z)$ может быть нацелен на максимальную автоинформативность новой системы показателей Z , т.е. на максимально точное воспроизведение всех исходных признаков $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ по сравнительно небольшому числу вспомогательных переменных $z^{(1)}, \dots, z^{(p')}$ ($p' \ll p$). В этом случае говорят о наилучшем автопрогнозе и обращаются к моделям и методам факторного анализа и его разновидностей (см. п. 13.2 и 13.3).

Сжатие массивов обрабатываемой и хранимой информации. Этот тип задач тесно связан с предыдущим и, в частности, требует в качестве одного из основных приемов решения построения экономной системы вспомогательных признаков, обладающих наивысшей автоинформативностью, т.е. свойством наилучшего автопрогноза (см. выше). В действительности при решении достаточно серьезных задач сжатия больших массивов информации (подобные задачи весьма актуальны и в плане необходимости минимизации емкостей носителей, на которых хранится архивная информация, и в плане экономии памяти ЭВМ при обработке текущей информации) используется сочетание методов классификации и снижения размерности. Методы классификации позволяют подчас перейти от массива, содержащего информацию по всем n статистически обследованным объектам, к соответствующей информации только по k эталонным образцам ($k \ll n$), где в качестве эталонных образцов берутся специальным образом отобранные наиболее типичные представители классов, полученных в результате операции разбиения исходного множества объектов на однородные группы. Методы же снижения размерности позволяют заменить исходную систему показателей $X = (x^{(1)}, \dots, x^{(p)})^T$ набором вспомогательных (наиболее автоинформативных) переменных $Z(X) = (z^{(1)}(X), \dots, z^{(p')}(X))^T$. Таким образом, размерность информационного массива понижается от $p \cdot n$ до $p' \cdot k$, т.е. во многие десятки раз, если учесть, что p' и k обычно на порядки меньше соответственно p и n .

Визуализация (наглядное представление) данных. При формировании рабочих гипотез, исходных допущений о геометрической и вероятностной природе совокупности анализируемых данных $X_1, X_2, \dots, X_n$

важно было бы суметь «подсмотреть», как эти данные-точки располагаются в анализируемом пространстве. В частности, уже на предварительной стадии исследования хотелось бы знать, например, распадается ли исследуемая совокупность точек на четко выраженные сгустки в этом пространстве, каково примерное число этих сгустков и т. д.? Но максимальная размерность «фактически осязаемого» пространства, как известно, равна трем. Поэтому, естественно, возникает проблема: нельзя ли спроецировать анализируемые многомерные данные из исходного пространства на прямую, на плоскость, в крайнем случае — в трехмерное пространство, но так, чтобы интересующие нас специфические особенности исследуемой совокупности (например, ее расслоенность на кластеры), если они присутствуют в исходном пространстве, сохранились бы и после проецирования. Следовательно, и здесь речь идет о снижении размерности анализируемого признакового пространства, но снижении, во-первых, подчиненном некоторым специальным критериям и, во-вторых, оговоренном условием, что размерность редуцированного пространства не должна превышать трех. Аппарат для решения подобных задач представлен в учебнике методом главных компонент и факторным анализом¹.

Построение условных координатных осей (многомерное шкалирование, латентно-структурный анализ). В данном типе задач снижение размерности понимается иначе, чем прежде. До сих пор речь шла о подчиненном некоторым специальным целям переходе от заданной координатной системы X (т. е. от исходных переменных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$) к новой координатной системе $Z(X)$, размерность которой p' существенно меньше размерности p и оси которой $0z^{(1)}, \dots, 0z^{(p')}$ конструируются с помощью соответствующих преобразований исходных признаков. Однако в данной постановке задачи *исходной координатной системы не существует вовсе*, а подлежащие статистическому анализу и моделированию данные представлены в *статическом* варианте (9.2), т. е. в виде матрицы (γ_{ij}) , $i, j = 1, 2, \dots, n$ — парных сравнений объектов. Ставится задача: для заданной, сравнительно невысокой, размерности p' определить вспомогательные условные координатные оси $0z^{(1)}, \dots, 0z^{(p')}$ и способ сопоставления каждому объекту O_i его координат $(z_i^{(1)}, \dots, z_i^{(p')})$ в этой системе таким образом, чтобы попарные отношения $\gamma_{ij}(Z)$ (например, попарные взаимные расстояния) между объектами, вычисленные на

¹ Наиболее эффективные современные методы для решения задач подобного типа объединяются в подходе, называемом «целенаправленным проецированием многомерных наблюдений». Однако описание этого подхода выходит за рамки целей данного учебника (см. [Айвазян С.А., Буштабер В.М. и др., 1989]).

базе этих условных координат, в определенном смысле минимально бы отличались от заданных величин γ_{ij} ($i, j = 1, 2, \dots, n$). В определенных условиях (в первую очередь в задачах педагогики, психологии, построения различных рейтингов² и т. п.) построенные таким образом условные переменные поддаются содержательной интерпретации и могут тогда рассматриваться в качестве латентных характеристик определенных свойств анализируемых объектов (такого типа задачи называют часто задачами латентно-структурного анализа). Снижение размерности происходит здесь в том смысле, что от исходного массива информации размерности $n \cdot n$ переходим к матрице типа «объект — свойство» размерности $p' \cdot n$, где $p' \ll n$. Аппарат для решения подобных задач состоит из методов так называемого *многомерного шкалирования* и представлен в п. 13.6.

9.1.3. Новые постановки задач и ослабление ограничительных условий в канонических математико-статистических и эконометрических моделях

Формализация (математическая постановка) реальных задач статистического анализа и моделирования экономических данных и на базе этого опыта выработка типовых математических постановок задач, выходящих за стеснительные рамки жестких канонических моделей, составляют, пожалуй, самый важный и самый трудный этап прикладного статистического исследования. Одновременно это направление деятельности в прикладной статистике является и самым неблагодарным, поскольку *de facto* оказался как бы «незаконнорожденным дитем» теории и практики статистического анализа данных. Искусство реалистического моделирования формально не предусмотрено ни в одном из разделов математико-статистической науки, а его развитие никак и ничем не стимулируется в рамках этой дисциплины.

Тем не менее за последние десятилетия прикладная статистика предложила и проанализировала ряд новых интересных типовых постановок математико-статистических задач, непосредственно стимулированных запросами практики (в том числе в социально-экономической сфере). В этой связи можно упомянуть об уже успевших стать каноническими постановках задач *обобщенного метода наименьших квадратов* (см. том 2, гл. 2),

² В подобных ситуациях элементы γ_{ij} матрицы (9.2) отражают обычно результаты попарных сравнений объектов O_i и O_j по анализируемому качеству. Например, численное значение γ_{ij} может отражать степень уверенности эксперта (или каким-то способом вычисленную вероятность) того, что объект O_i лучше объекта O_j .

о различных подходах к построению *устойчивых статистических выводов*, о развитии методов *целенаправленного проецирования и томографии* в связи с исследованием геометрической и вероятностной природы анализируемых многомерных данных (9.1) (см., например, [Айвазян С. А., Енюков И. С., Мешалкин Л. Д., Бухштабер В. М., 1989]), а в последние годы — о формулировке и исследовании таких моделей анализа временных рядов, как *модель коинтеграции, ARCH- и GARCH-модели* (см. гл. 3 тома 2), *модель многомерной развертки временного ряда* [Бухштабер В. М.] и др.

9.2. Основные этапы прикладного статистического анализа

Для пояснения роли и места основных приемов статистического моделирования и методов первичной статистической обработки исходных данных удобно разложить общую логическую схему статистического анализа на *основные этапы исследования*. Подобное разложение носит, конечно, условный характер. В частности, оно не означает, что этапы осуществляются в строгой хронологической последовательности один за другим. Более того, многие из этапов (например, этапы 4, 5 и 6) находятся в плане хронологическом в соотношении итерационного взаимодействия: результаты реализации более поздних этапов могут содержать выводы о необходимости повторной «прогонки» (с учетом новой информации) предыдущих этапов.

Этап 1: исходный (предварительный) анализ исследуемой реальной системы. В результате этого анализа определяются: а) основные цели исследования на неформализованном, содержательном уровне; б) совокупность единиц, представляющая предмет статистического исследования; в) перечень $(x^{(1)}, x^{(2)}, \dots, x^{(p)})$ отобранных из представленного специалистами априорного набора показателей, характеризующих состояние (поведение) каждого из обследуемых объектов, который предполагается использовать в данном исследовании; г) степень формализации соответствующих записей при сборе данных; д) общее время и трудозатраты, отведенные на планируемые работы, и коррелированные с ними временная протяженность и объем необходимого статистического обследования; е) моменты, требующие предварительной проверки перед составлением детального плана исследования (например, не всегда априори ясна возможность идентификации единиц наблюдения); ж) формализованная постановка задачи, по возможности включающая вероятностную модель изучаемого явления и природу статистических выводов, к которым должен

(или может) прийти исследователь в результате переработки массива исходных данных; 3) формы, используемые для сбора первичной информации и для введения ее в вычислительное устройство.

По затратам сил наиболее квалифицированного персонала, участвующего в работе, трудоемкость первого этапа работы весьма значительна и бывает даже сравнима с суммарной трудоемкостью всех остальных этапов при условии, что обработка проводится с помощью подходящего пакета программ¹. Поэтому максимального развития заслуживают методы *машинного ассистирования* в проведении этой части работы. Оно может заключаться в подсказке (с одновременной оценкой) форм документации для сбора первичной информации, медицинских приложений), в автоматизированном режиме выбора подходящих моделей, в ведении тезауруса исследования и т. п.

Этап 2: составление детального плана сбора исходной статистической информации. При составлении этого плана необходимо по возможности учитывать полную схему дальнейшего статистического анализа, о чем часто забывают. Априорное представление о том, как и для чего данные будут анализироваться, может оказать существенное влияние на их сбор. При планировании особого внимания заслуживают случаи, когда: а) используется аппарат теории выборочных обследований, т. е. определяется, какой должна быть выборка — случайной, пропорциональной, расслоченной и т. п.; б) хотя бы для части входных переменных эксперимент носит *активный* характер, т. е. переменные допускают фиксацию в каждом конкретном наблюдении на определенном уровне, и выбор плана обследования осуществляется с привлечением методов планирования экспериментов. В некоторых руководствах по статистике этот этап называют этапом *«организационно-методической подготовки»*. Как уже было упомянуто в п. 6.1, вопросы разработки методологии определения априорной системы показателей, характеризующих исследуемый объект или процесс, вынесены за рамки описываемых здесь этапов и должны быть отнесены к области предметной (в нашем случае — экономической) статистики.

Этап 3: сбор исходных статистических данных и их ввод в вычислительное устройство. Одновременно в вычислительное устройство вносятся полные и краткие (для автоматизированного воспроизводства в таблицах) определения используемых терминов. В программном обеспече-

¹ В некоторых специальных статистических исследованиях социально-экономического и других профилей, характеризующихся большими затратами времени и средств на сбор исходных статистических данных, сформулированный тезис остается справедливым лишь при условии исключения этапа 3 из суммарной трудоемкости всех остальных этапов.

нии должны быть предусмотрены специальные меры, исключаящие или резко уменьшающие возможность появления расчетов не с тем подмножеством данных или не для той подгруппы объектов.

Таким образом, независимо от того, производится ли исследователем выбор метода и плана статистического обследования или он уже располагал результатами так называемого *пассивного эксперимента*, к моменту определения основного инструментария статистического исследования исследователь в общем случае располагает в качестве массива исходных статистических данных матрицами наблюдений вида (9.1), (9.1') или (9.2).

Этап 4: первичная статистическая обработка данных. В ходе первичной статистической обработки данных обычно решаются следующие задачи: а) отображение переменных, описанных текстом, в номинальную (с предписанным числом градаций) или ординальную (порядковую) шкалу; б) статистическое описание исходных совокупностей с определением пределов варьирования переменных; в) анализ резко выделяющихся наблюдений; г) восстановление пропущенных наблюдений; д) проверка статистической независимости последовательности наблюдений, составляющих массив исходных данных; е) унификация типов переменных, когда с помощью различных приемов добиваются унифицированной записи всех переменных; ж) экспериментальный анализ закона распределения исследуемой генеральной совокупности и параметризация сведений о природе изучаемых распределений (иногда этот этап называют процессом составления *сводки и группировки*). Кроме того, этап 4 включает в себя вычислительную реализацию решения следующих вопросов: учет размерности и алгоритмической сложности задачи и одновременно возможностей используемого вычислительного средства; формулировку задачи на входном языке используемого программного обеспечения и т. п. (см. подробнее об этом в описании этапа 6).

Остановимся на некоторых из затронутых вопросов подробнее.

Анализ резко выделяющихся наблюдений. Часто даже беглый предварительный просмотр (визуальный или автоматизированный) исходных данных (9.1) или (9.2) может вызвать у исследователя сомнения в истинности (или правомерности) отдельных наблюдений, слишком резко выделяющихся на общем фоне. В этих случаях возникает вопрос: вправе ли мы объяснить обнаруженные резкие отклонения в исходных данных (аномальные выбросы) лишь обычными случайными колебаниями выборки (которые обусловлены природой анализируемой генеральной совокупности) или здесь дело в существенных искажениях стандартных условий сбора статистических данных, а возможно, и в прямых ошибках регистрации (записи)? В последних двух случаях «подозрительные» наблюдения,

очевидно, следует исключить из дальнейшего рассмотрения.

Единственным абсолютно надежным способом решения вопроса об исключении резко выделяющихся результатов наблюдений является тщательное рассмотрение условий, при которых эти наблюдения регистрировались. Однако во многих случаях проведение такого содержательного анализа объективно затруднительно или принципиально невозможно. Тогда необходимо обратиться к соответствующим формальным (статистическим) методам. Общая логическая схема этих методов следующая: отправляясь от исходных допущений о природе анализируемой совокупности данных, исследователь задается функцией

$$\rho(X^*; X_1, X_2, \dots, X_n) \quad (9.3)$$

от всех имеющихся наблюдений, характеризующей степень аномальности (меру удаленности от основной массы наблюдений) «подозрительного» наблюдения X^* , а затем подставляет в (9.3) реальные значения наблюдений и сравнивает величину с некоторым пороговым значением ρ_0^1 ; если $\rho > \rho_0$, то подозрительное наблюдение X^* или полностью исключается из дальнейшего рассмотрения, или его вклад уменьшается с помощью весовой функции, убывающей по мере роста степени аномальности наблюдений (см. п. 7.5.3).

С одним из вариантов методов анализа резко выделяющихся наблюдений читатель познакомился в п. 8.6.

Восстановление пропущенных (стертых) наблюдений. В матрицах исходных статистических данных (9.1) или (9.2) по разным причинам (в том числе и в результате исключения резко выделяющихся наблюдений) могут быть *пропуски* отдельных элементов или каких-то частей строк или столбцов. Исключать по этой причине из дальнейшего рассмотрения весь объект (строку, в которой обнаружены пропуски) или признак (столбец, в котором обнаружены пропуски) слишком расточительно с точки зрения потери полезной информации. Поэтому возникает задача наилучшего в некотором смысле восстановления пропущенных (стертых) данных. Конкретизация критерия качества восстановления стертых данных производится в зависимости от характера последующей обработки исходных данных, т. е. в зависимости от окончательных целей исследования (см., например, [Айвазян С. А., Енюков И. С., Мешалкин Л. Д., 1983]).

¹ В вероятностной постановке задачи пороговое значение ρ_0 определяется из стандартных статистических таблиц с учетом знания закона распределения статистики ρ в предположении необоснованности «подозрений» относительно наблюдения X^* . В других случаях ρ_0 определяется из содержательных соображений.

Проверка однородности нескольких порций исходных данных. Объективные условия сбора исходных статистических данных, особенно в ситуациях пассивного эксперимента, могут быть такими, что общая $(p \times n)$ -матрица наблюдений (см. (9.1)) получается составлением нескольких $(p \times n_1)$ -, $(p \times n_2)$ -, ..., $(p \times n_k)$ -матриц (частных) наблюдений ($n_1 + n_2 + \dots + n_k = n$), где каждая из частных матриц задает порцию исходных данных, относящихся к некоторой подсовокупности, состоящей из n_j объектов. При этом процессы (моменты) обследования этих совокупностей могут быть разделены в пространстве (во времени).

Очевидно, перед тем как подвергнуть исходные данные основной статистической обработке (т.е. применять к ним те или иные методы прикладного статистического анализа, выбор которых обусловлен конечными целями исследования), исследователь должен ответить на вопрос: правомерно ли объединение имеющихся в его распоряжении порций (выборок) в один общий массив или же каждая из порций имеет свою специфику, следовательно, и обрабатывать их надо по отдельности? В рамках математико-статистических моделей этот вопрос сводится к выяснению (с помощью соответствующих статистических критериев), можно ли считать эти порции данных различными выборками из *одной и той же генеральной совокупности* (см., например, пп. 8.1.2, 8.6.3, 8.6.4).

Проверка статистической независимости последовательности наблюдений, составляющих массив исходных данных. Применение многих статистических методов является правомерным лишь в ситуациях, когда справедливо допущение о статистической независимости обрабатываемого ряда наблюдений $X_1, X_2, \dots, X_n$. Этот же вопрос возникает и применительно к рядам $\{X_i(t_1), \dots, X_i(t_N)\}$. Поэтому, перед тем как подвергнуть имеющиеся результаты наблюдения основной статистической обработке, необходимо выяснить (с помощью соответствующих статистических критериев (см. п. 3.1 тома 2)), являются ли они статистически независимыми или их следует рассматривать как последовательности взаимозависимых величин.

Унификация типа переменных. Одна из сложностей автоматизированного анализа информации заключается в том, что среди компонент $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ анализируемого многомерного признака могут быть показатели трех разных типов: *количественные*, *качественные* (порядковые, ординальные) и *классификационные* (номинальные). Их определение и сущность приведены в п. 2.3.

В связи с этим возникает вопрос унификации записи единичного наблюдения, снятого с объекта i . В соответствии с одним из вариантов решения этого вопроса i -е многомерное наблюдение в унифицированной записи

представляется вектор-столбцом размерности $m_1 + m_2 + \dots + m_p$, где m_k — число градаций (интервалов группирования, уровней качества или однородных групп) признака $x^{(k)}$, причем компонентами этого вектор-столбца могут быть только нули или единицы. При таком подходе к достижению единообразия записи наблюдений многомерного признака смешанной природы мы вынуждены мириться, во-первых, с элементами субъективизма в выборе способов разбиения диапазонов изменения анализируемых количественных признаков на интервалы группирования и, во-вторых, с определенной потерей информативности исходных данных, связанной с переходом от *индивидуальных* к *группированным* значениям по количественным переменным.

В качестве альтернативного подхода к способу унификации записи исходных данных может быть использована идея, прямо противоположная той, на основании которой построен только что описанный прием. В частности, руководствуясь некоторыми дополнительными соображениями (и допущениями), исследователь пытается преобразовать качественные и классификационные переменные в количественные, используя процесс так называемой «оцифровки», или шкалирования, неколичественных переменных, а также некоторые специальные модели (см. [Айвазян С. А., Енюков И. С., Мешалкин Л. Д., 1983, п. 10.2]).

Экспериментальный анализ закона распределения исследуемой генеральной совокупности и вопрос ее подходящей параметризации. Эта часть предварительной статистической обработки исходного массива данных, представленных в виде (9.1), включает в себя вычисление основных числовых характеристик распределения: среднего значения, дисперсии, коэффициентов асимметрии и эксцесса, а в многомерном случае — и элементов выборочной ковариационной матрицы. Кроме того, исследователь проводит численный и графический анализ одномерных законов распределения рассматриваемых показателей, заключающийся в построении соответствующих полигонов частот, гистограмм, эмпирических функций распределения. Результаты этого экспериментального анализа, дополненные априорными сведениями о природе анализируемой генеральной совокупности, зачастую оказываются достаточными для формулировки одной или нескольких конкурирующих гипотез об общем (параметрическом) виде закона распределения вероятностей, задающего эту генеральную совокупность. Не следует пренебрегать такой возможностью, поскольку знание общего вида вероятностного распределения в исследуемой генеральной совокупности позволяет сделать наилучший выбор метода статистического оценивания параметров этого распределения, а также метода последующей основной статистической обработки масси-

ва исходных данных (из набора конкурирующих методов). Как известно, выяснение непротиворечивости высказанной исследователем гипотезы об общем виде распределения анализируемых наблюдений с природой и спецификой имеющихся в распоряжении исследователя конкретных исходных данных осуществляется с помощью тех или иных статистических *критериев согласия* (см. пп. 8.1.1 и 8.6.1).

Этап 5: составление детального плана вычислительного анализа материала. Этап начинается с составления справки по собранному материалу и результатам предварительного анализа. Определяются основные группы, для которых будет проводиться дальнейший анализ. Пополняется и уточняется тезаурус содержательных понятий. Четко описывается блок-схема анализа с указанием привлекаемых методов. *Формулируется оптимизационный критерий*, на основании которого выбирается один из альтернативных методов (или одно из альтернативных семейств методов) основной статистической обработки исходных данных.

Этап 6: вычислительная реализация основной части статистической обработки данных. Основная забота исследователя на этом этапе — эффективное управление вычислительным процессом путем формулировки задачи обработки и описания данных на входном языке используемого программного обеспечения. Учитываются размерность задачи, алгоритмическая сложность вычислительного процесса, возможности используемого вычислительного средства (длина слова, быстродействие, объем оперативной памяти, организация базы данных и т. п.) и, наконец, особенности данных (степень обусловленности используемых при реализации линейных процедур матриц, надежность априорных оценок параметров и т. п.).

Этап 7: подведение итогов исследования. Этап начинается с построения формального статистического отчета о проведенном исследовании. При интерпретации результатов применения статистических процедур (оценка параметров, проверка гипотез, отображения в пространство меньшей размерности, классификация и т. п.) учитывается как место этих процедур в блок-схеме анализа, так и соотношение объемов используемых выборок, размерности пространства наблюдений, числа и значений параметров. Теоретически эти вопросы, несмотря на их крайнюю актуальность, разработаны довольно мало.

Затем результаты исследования, его основные выводы формулируются в содержательных терминах. Если исследование проводилось в рамках математико-статистических методов и моделей, то его выводы формулируются в терминах оценок неизвестных параметров анализируемой системы или в виде ответа на вопрос о справедливости проверяемой гипотезы и

сопровождаются гарантируемыми количественными оценками степени их достоверности. Если же исследование осуществлялось средствами анализа данных (т. е. в рамках логико-алгебраического подхода), то его выводы не претендуют на вероятностную интерпретацию.

В заключение проверяется, в какой мере достигнуты намеченные на этапе 1 содержательные цели работы, и если достигнуты не все из них, то объясняется, почему. Работа завершается содержательной формулировкой новых задач, вытекающих из проведенного исследования.

В некоторых руководствах по статистике этапы 5, 6 и 7 объединены в одном этапе, названном «*Обработка и анализ*».

Резюмируя описание общей логической схемы статистического анализа исходных данных, отметим, что основные приемы статистического моделирования и методы первичной статистической обработки являются главными в ходе реализации важнейших этапов 1, 4 и 7, а также по мере необходимости могут привлекаться при реализации этапов 3, 5 и 6.

ВЫВОДЫ

1. *Прикладная статистика* заимствует часть методов статистического анализа и моделирования у математической статистики и имеет пересечения в задачах и целях с эконометрикой. Однако основными характеристными чертами, отличающими прикладную статистику от этих двух дисциплин, являются следующие:

- основу методологии разработки инструментария прикладной статистики составляет *подход, позволяющий в рамках единой унифицированной логической схемы синтезировать две автономно возникших и какое-то время взаимонезависимо развивавшихся концепции анализа данных и моделирования — вероятностную и логико-алгебраическую*; в результате прикладная статистика объединила в себе как прикладные вероятностно-статистические методы многомерного статистического анализа, так и логико-алгебраические методы анализа данных, не использующие в своих построениях вероятностных рассуждений;
- главный исследовательский интерес прикладной статистики сосредоточен на трех проблемах: (а) *статистическом исследовании зависимостей между анализируемыми показателями*; (б) *статистических методах классификации объектов и признаков*; (в) *методах снижения размерности исследуемого признакового пространства с целью лаконичного объяснения природы анализируемых многомерных данных*;

- прикладная статистика, в отличие от математической статистики и в какой-то мере эконометрики относится к весьма динамичным по своему содержанию и постановкам задач научным дисциплинам: *она постоянно нацелена на выработку новых (стимулированных запросами практики) постановок задач и на достижение результатов, позволяющих ослабить ограничительные условия в канонических математико-статистических и эконометрических моделях.*

2. Двумя наиболее распространенными формами представления исходных статистических данных, поступающих «на вход» прикладного статистического анализа и моделирования, являются *матрицы «объект-свойство»* вида (9.1) и *матрицы парных сравнений* (объектов или признаков) вида (9.2). Первая из них содержит результаты измерений анализируемого набора признаков $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^T$ на статистически обследуемом множестве объектов $O_1, O_2, \dots, O_n$, а вторая (чаще всего получаемая с помощью специальных анкет и экспертных опросов) состоит из характеристик попарных сравнений статистически обследуемых объектов по некоторому анализируемому свойству.

3. *Проблема статистического исследования зависимостей* состоит, в самой общей формулировке, в выявлении и описании парных и множественных статистических связей (их смысла, степени тесноты, формы и т.п.), существующих между компонентами анализируемого многомерного признака $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^T$. Причем анализ этих связей производится на основании исходных статистических данных вида (9.1) или (9.2). Описанию задач и инструментария, относящихся к этой проблеме, посвящены гл. 10, 11, а также гл. 2, 3 и 4 в томе 2.

4. *Проблема классификации объектов* (в общей, нестрогой, постановке) заключается в том, чтобы всю анализируемую совокупность объектов $O_1, O_2, \dots, O_n$, статистически представленную в виде матриц (9.1) или (9.2), разбить на сравнительно небольшое число (заранее известное или нет) однородных, в определенном смысле, групп или классов. Описанию задач и инструментария, относящихся к этой проблеме, посвящена гл. 12.

5. *Проблема снижения размерности анализируемого признакового пространства* состоит в таком переходе от исходного набора p признаков $(x^{(1)}, x^{(2)}, \dots, x^{(p)})$ к вспомогательному набору гораздо меньшего числа p' признаков-детерминант $z^{(1)}, z^{(2)}, \dots, z^{(p')}$, при котором либо наблюдаемые значения самих исходных признаков (т.е. данные матрицы (9.1)), либо некоторые интересующие нас свойства или характеристики, определяемые исходными признаками, могут быть наиболее точно (в определенном смысле) оценены по совокупности значений вспомогательных признаков $z^{(1)}, z^{(2)}, \dots, z^{(p')}$. При этом упомянутые признаки-детерминанты могут

выбираться из числа исходных, а могут конструироваться в качестве некоторых функций от исходных признаков. К методам снижения размерности обращаются при решении ряда конкретных типовых прикладных задач статистического анализа и моделирования, среди которых:

- отбор наиболее информативных показателей в задачах классификации (п. 12.3 и 12.4), в моделях регрессии (п. 2.6 в томе 2) и при построении интегральной латентной характеристики («агрегатного скалярного индикатора») изучаемого свойства (п. 13.4);

- сжатие больших массивов обрабатываемой и хранимой информации;
- визуализация (наглядное представление) многомерных статистических данных;

- построение условных координатных осей для описания «состояний» объектов, статистически заданных матрицей парных сравнений (9.2).

6. Общая логическая схема статистического анализа какого-либо процесса или явления может быть условно разложена на следующие семь этапов:

- 1) предварительный анализ исследуемой реальной системы;
- 2) составление детального плана сбора исходной статистической информации;
- 3) сбор исходных статистических данных;
- 4) преданализ или первичная статистическая обработка данных;
- 5) составление плана моделирования и вычислительного анализа материала;
- 6) вычислительная реализация методов статистического анализа и моделирования;
- 7) интерпретация результатов и подведение итогов исследования.

В рамках этой схемы удобно комментировать роль и место основных разделов и методов прикладной статистики.

ГЛАВА 10. СТАТИСТИЧЕСКОЕ ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ (ОСНОВНЫЕ ПОНЯТИЯ И ПОСТАНОВКИ ЗАДАЧ)

10.1. Общая формулировка проблемы, пример

Любой закон природы или общественного развития может быть выражен в конечном счете в виде описания характера или структуры взаимосвязей (зависимостей), существующих между изучаемыми явлениями или показателями (переменными величинами или просто *переменными*). Если эти зависимости: а) *стохастичны* по своей природе, т. е. позволяют устанавливать лишь вероятностные логические соотношения между изучаемыми событиями A и B , а именно соотношения типа «из факта осуществления события A следует, что событие B должно произойти, но не обязательно, а лишь с некоторой (как правило, близкой к единице) вероятностью P »; б) выявляются на основании *статистического наблюдения* за анализируемыми событиями или переменными, осуществляемого по выборке из интересующей нас генеральной совокупности (см. гл. 6), то мы оказываемся в рамках проблемы *статистического исследования зависимостей*. Соответствующий математический аппарат, будучи таким образом нацеленным в первую очередь на решение основной проблемы естествознания: как по отдельным, частным наблюдениям выявить и описать интересующую нас общую закономерность, — занимает, бесспорно, центральное место во всем прикладном математическом анализе.

Перед тем как перейти к формулировке общей и частных задач статистического исследования зависимостей, условимся описывать функционирование изучаемого реального объекта (системы, процесса, явления) набором переменных (рис. 10.1), среди которых:

$x^{(1)}, x^{(2)}, \dots, x^{(p)}$ — так называемые «входные» *переменные*, опи-

сывающие *условия функционирования* (часть из них, как правило, поддается регулированию или частичному управлению); в соответствующих математических моделях их называют независимыми, факторами-аргументами, предикторными (или просто предикторами, т. е. предсказателями), экзогенными, объясняющими (в книге мы будем использовать в основном два последних термина);

$y^{(1)}, y^{(2)}, \dots, y^{(m)}$ — *выходные переменные*, характеризующие *поведение или результат* (эффективность) *функционирования*; в математических моделях их называют зависимыми, откликами, эндогенными, результирующими или объясняемыми (в книге используются в основном три последних термина);

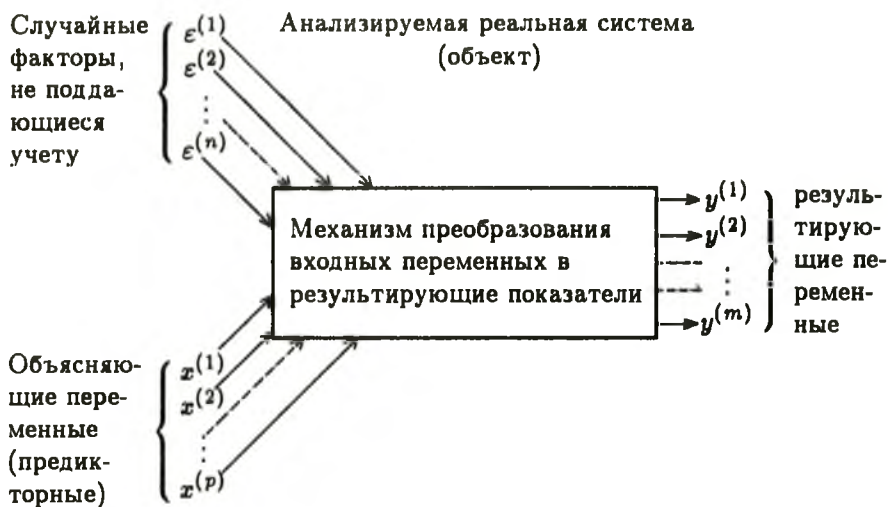


Рис. 10.1. Общая схема взаимодействия переменных при статистическом исследовании зависимостей

$\varepsilon^{(1)}, \varepsilon^{(2)}, \dots, \varepsilon^{(m)}$ — латентные (т. е. скрытые, не поддающиеся непосредственному измерению) *случайные «остаточные» компоненты*, отражающие влияние (соответственно на $y^{(1)}, y^{(2)}, \dots, y^{(m)}$) не учтенных «на входе» факторов, а также случайные ошибки в измерении анализируемых показателей (в математических моделях мы их, как правило, будем именовать просто «остатками»).

Тогда общая задача статистического исследования зависимостей (в терминах изучаемых показателей) может быть сформулирована следующим образом:

по результатам n измерений

$$\{(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)}; y_i^{(1)}, y_i^{(2)}, \dots, y_i^{(m)})\}_{i=1,2,\dots,n} \quad (10.1)$$

исследуемых переменных на объектах (системат, процессах) анализируемой совокупности построить такую (векторнозначную) функцию

$$f(x^{(1)}, x^{(2)}, \dots, x^{(p)}) = \begin{pmatrix} f^{(1)} & (x^{(1)}, \dots, x^{(p)}) \\ f^{(2)} & (x^{(1)}, \dots, x^{(p)}) \\ \vdots & \vdots \\ f^{(m)} & (x^{(1)}, \dots, x^{(p)}) \end{pmatrix}, \quad (10.2)$$

которая позволила бы наилучшим (в определенном смысле) образом восстанавливать значения результирующих (прогнозируемых) переменных $Y = (y^{(1)}, y^{(2)}, \dots, y^{(m)})^T$ по заданным значениям объясняющих (экзогенных) переменных $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^T$.

Данная формулировка задачи нуждается в уточнениях. В частности, прежде всего мы должны ответить на следующие вопросы:

а) каково математическое выражение (или структура модели) истинной зависимости между Y и X , записанное в терминах $Y, X, f(X)$ и $\varepsilon = (\varepsilon^{(1)}, \varepsilon^{(2)}, \dots, \varepsilon^{(m)})^T$?

б) в соответствии с каким критерием качества аппроксимации значений Y с помощью функции $f(X)$ мы будем определять наилучший способ восстановления значений результирующих показателей по заданным значениям объясняющих переменных?

в) с какой именно прикладной целью мы проводим все наше исследование, т. е. для решения каких конкретных задач мы собираемся использовать построенную в результате исследования функцию $f(X)$?

Прежде чем обсуждать эти вопросы, рассмотрим пример.

Пример 10.1. Анализируется «поведение» двумерной случайной величины (ξ, η) , где ξ (ден.ед.) — среднедушевой доход и η (ден.ед.) — среднедушевые денежные сбережения в семье, случайно извлеченной из рассматриваемой совокупности семей, однородной по своему потребительскому поведению. В табл. 10.1 и на рис. 10.2 представлены исходные статистические данные вида (10.1), характеризующие среднедушевые величины дохода (x_i) и денежных сбережений (y_i) за определенный отрезок времени, а именно за месяц, в каждой $(i$ -й, $i = 1, 2, \dots, n)$ обследованной семье рассматриваемой совокупности семей (в данном условном примере объем n статистически обследованной совокупности семей равнялся 40). В этом примере имелась возможность при отборе исходных данных (выборки) контролировать значения экзогенной переменной ξ , что позволило, в

частности, разбить статистически обследованные семьи на четыре равные по объему группы по доходам.

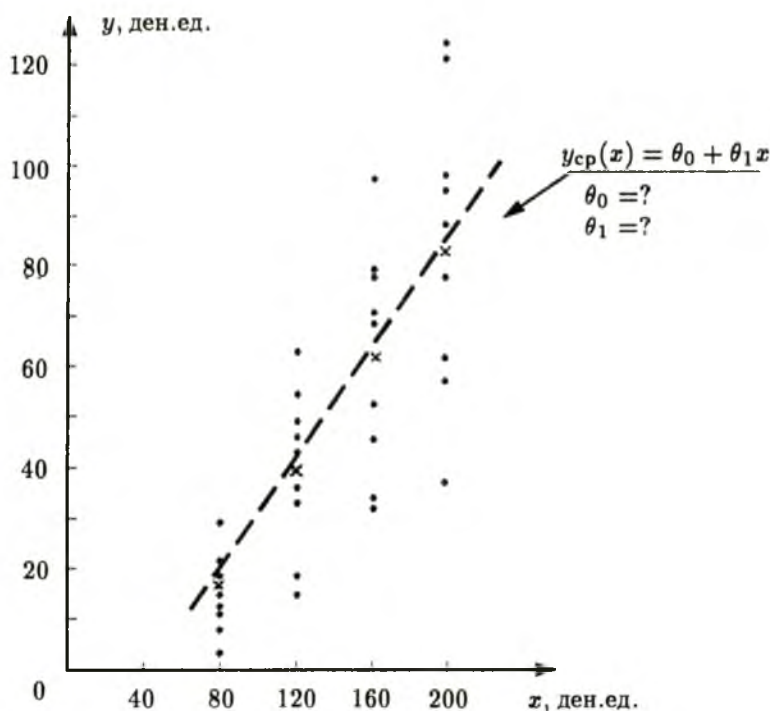


Рис. 10.2. Графическое представление результатов обследования 40 семей по их среднему доходу (x_i) и среднему денежному сбережению (y_i)

Мы видим, что даже в пределах каждой из этих групп величины среднечеловеческих сбережений семей подвержены некоторому неконтролируемому разбросу, обусловленному влиянием множества не поддающихся строгому учету и контролю факторов (т.е. налицо упомянутый выше *стохастический характер* зависимости между x и y). Однако это еще не значит, что расположение точек (x_i, y_i) , являющихся геометрическим изображением результатов обследования семей по доходу и сбережениям, должно быть совершенно хаотичным и не должно обнаруживать некоторой *вполне определенной тенденции*, характеризующей зависимость денежных сбережений в семье (η) от ее среднечеловеческого дохода (ξ). При исследовании подобных зависимостей встают следующие основные вопросы (в скобках после вопроса указываются главы и пункты настоящей книги, ему посвя-

щенные).

1. Как исходя из конкретных прикладных целей исследования определить смысл, в котором понимается исследуемая зависимость? (п. 10.2 и 10.5).

2. Имеется ли вообще какая-либо связь между исследуемыми переменными (а в случае многих переменных какова структура этих связей?) и как измерить тесноту этой связи? (гл. 11).

3. Каков *общий* математический вид искомой связи между η и ξ , т.е. как определяется *общая структура* соответствующей математической модели? (п. 10.7).

4. Как, отправляясь от принятой общей структуры модели, провести необходимую вычислительную обработку исходных данных (10.1) с целью получения *конкретного вида* зависимости η от ξ , что позволит в данном случае производить количественную оценку неизвестных денежных сбережений семьи по заданной величине ее среднедушевого дохода? (гл. 2, 3, 4 в томе 2).

5. Поскольку наши выводы основаны на обработке *ограниченного ряда* наблюдений, то их количественные характеристики, естественно, подвержены (при повторениях соответствующих выборочных обследований) некоторому случайному разбросу. Как оценить *степень точности наших выводов*? (гл. 2, 3, 4 в томе 2).

6. Как решать сформулированные выше вопросы в ситуациях, когда среди объясняющих (предикторных) переменных могут быть и неколичественные? (ответы на 2-й вопрос — в гл. 11).

Вернемся к нашему примеру и попробуем ответить на некоторые из поставленных здесь вопросов, в том числе на принципиальные вопросы а), б) и в), ответы на которые позволяют уточнить общую формулировку задачи статистического исследования зависимостей, данную выше.

Начнем «с конца», т.е. с уточнения *конечных прикладных исследований* (см. вопросы 1, а также а) и в)). Известно, что из двух анализируемых характеристик материальной состоятельности семьи характеристика денежных сбережений (η) относится к категории статистически труднодоступных: содержащиеся в ежегодных и единовременных выборочных семейных бюджетных обследованиях Госкомстата РФ сведения о сбережениях, как правило, ненадежны и непредставительны. Поэтому главной конечной целью нашего исследования (опирающегося, как мы будем всегда предполагать, на достоверную и репрезентативную выборку исходных данных) является возможность восстановления (прогноза):

• удельной (т.е. в расчете на одного члена семьи за определенный отрезок времени) величины денежных сбережений в *конкретной семье* ($y(x)$)

по заданному значению ее среднедушевого дохода x ;

• удельной величины *средних* денежных сбережений ($y_{\text{ср}}(x)$) в семьях данной группы x по доходам.

Этой цели мы сможем достигнуть, если сумеем математически описать закономерность изменения условных теоретических средних значений $y_{\text{ср}}(x) = E(\eta | \xi = x)$ в зависимости от x , а также изучить характер случайного разброса денежных сбережений $y(x)$ отдельных семей данной группы x по доходам относительно своего среднего значения $y_{\text{ср}}(x)$ (при любом интересующем нас значении среднедушевого дохода x). Это естественным образом приводит нас к необходимости рассмотрения математической модели вида

$$\eta = f(x) + \varepsilon, \quad (10.3)$$

в которой остаточная компонента ε отражает случайное отклонение денежных сбережений наугад выбранной отдельной семьи с доходом $\xi = x$ от среднего значения $y_{\text{ср}}(x) = E(\eta | \xi = x)$ этих сбережений, подсчитанного по всем семьям данной группы по доходам, а функция $f(x)$ описывает характер изменения условного среднего $y_{\text{ср}}(x)$ (при $\xi = x$) в зависимости от изменения x , если дополнительно прийти к соглашению, что характер случайного разброса величин $y(x) = (\eta | \xi = x)$ относительно своих средних $y_{\text{ср}}(x)$ таков, что $E(\varepsilon | \xi = x) = 0$ при всех x .

Таким образом, из (10.3) мы непосредственно получаем

$$y_{\text{ср}} = E(\eta | \xi = x) = f(x). \quad (10.4)$$

Чтобы покончить с вопросами 1, а) и в), остается уточнить общую структуру модели, т. е. определить, в каком классе F функций $f(x)$ мы будем производить аппроксимацию искомой зависимости $y_{\text{ср}}(x)$.

В нашем случае, учитывая однородный (по характеру потребительского поведения) состав исследуемой совокупности семей, естественно исходить из гипотезы об одинаковой (в среднем) склонности семей к сбережениям, выражающейся, в частности, в том, что все семьи начиная с некоторого «порогового» уровня дохода склонны отделять в сбережения в среднем одинаковую долю дохода. Математически, как легко понять, это выразится в виде

$$y_{\text{ср}}(x) = \theta_0 + \theta_1 x, \quad (10.5)$$

где θ_0 и θ_1 — некоторые константы (*неизвестные параметры модели*). Так что

$$F = \{\theta_0 + \theta_1 x\}, \quad (10.6)$$

где под $\{f(x; \Theta)\}$ понимается семейство всех тех функций $f(x; \Theta)$, которые могут быть получены при подстановке вместо Θ ее различных

конкретных значений (Θ — векторный параметр; в нашем случае $\Theta = (\theta_0, \theta_1)^T$).

Таблица 10.1

Средне- душевой доход (ден.ед.)	Средне- душевые сбереже- ния (ден.ед.)	Средние сбережения для семей данной группы (ден.ед.)	Оценка среднеквадра- тического отклонения $\bar{s}$ и коэффициента вариа- ции $\hat{V}$ сбережений для семей данной доход- ной группы
1	2	3	4
$x_1 = x_2 =$ $\dots = x_{10} =$ $= x_1^0 = 80$	$y_1 = 15,2$ $y_2 = 10,7$ $y_3 = 18,5$ $y_4 = 14,9$ $y_5 = 24,1$ $y_6 = 10,3$ $y_7 = 14,2$ $y_8 = 31,0$ $y_9 = 20,4$ $y_{10} = 20,0$	$\bar{y}(x_1^0) =$ $= \frac{1}{10} \sum_{i=1}^{10} y_i = 17,9$	$\bar{s}(x_1^0) = \left[\frac{1}{9} \sum_{i=1}^{10} (y_i - \bar{y}(x_1^0))^2 \right]^{\frac{1}{2}} = 6,4$ $\hat{V}(x_1^0) = 36\%$
$x_{11} = x_{12} =$ $\dots = x_{20} =$ $= x_2^0 = 120$	$y_{11} = 70,1$ $y_{12} = 35,0$ $y_{13} = 43,0$ $y_{14} = 29,0$ $y_{15} = 17,0$ $y_{16} = 48,2$ $y_{17} = 18,9$ $y_{18} = 53,0$ $y_{19} = 39,4$ $y_{20} = 46,2$	$\bar{y}(x_2^0) =$ $= \frac{1}{10} \sum_{i=11}^{20} y_i = 40,0$	$\bar{s}(x_2^0) = \left[\frac{1}{9} \sum_{i=11}^{20} (y_i - \bar{y}(x_2^0))^2 \right]^{\frac{1}{2}} = 16,0$ $\hat{V}(x_2^0) = 40\%$

Продолжение таблицы 10.1

1	2	3	4
$x_{21} = x_{22} =$ $\dots = x_{30} =$ $= x_3^0 = 160$	$y_{21} = 49,6$ $y_{22} = 69,4$ $y_{23} = 77,8$ $y_{24} = 43,0$ $y_{25} = 31,8$ $y_{26} = 62,6$ $y_{27} = 100,2$ $y_{28} = 68,8$ $y_{29} = 78,0$ $y_{30} = 29,6$	$\bar{y}(x_3^0) =$ $= \frac{1}{10} \sum_{i=21}^{30} y_i = 61,1$	$\bar{s}(x_3^0) = \left[\frac{1}{9} \sum_{i=21}^{30} (y_i - \bar{y}(x_3^0))^2 \right]^{\frac{1}{2}} = 22,6$ $\hat{V}(x_3^0) = 37\%$
$x_{31} = x_{32} =$ $\dots = x_{40} =$ $= x_4^0 = 200$	$y_{31} = 125,5$ $y_{32} = 88,3$ $y_{33} = 62,0$ $y_{34} = 58,8$ $y_{35} = 84,0$ $y_{36} = 79,0$ $y_{37} = 95,5$ $y_{38} = 120,8$ $y_{39} = 98,1$ $y_{40} = 29,7$	$\bar{y}(x_4^0) =$ $= \frac{1}{10} \sum_{i=31}^{40} y_i = 84,2$	$\bar{s}(x_4^0) = \left[\frac{1}{9} \sum_{i=31}^{40} (y_i - \bar{y}(x_4^0))^2 \right]^{\frac{1}{2}} = 28,9$ $\hat{V}(x_4^0) = 34\%$

Такой выбор «класса допустимых решений» $F = \{f(x)\}$ подтверждается и характером расположения совокупности точек, являющихся геометрическим изображением исходных данных в нашем примере (см. на рис. 10.2 расположение «крестиков», ординаты которых определяются экспериментально подсчитанными, т. е. вычисленными на основании имеющихся выборочных данных, условными средними $\bar{y}(x_i^0), i = 1, 2, 3, 4$)¹.

¹ Обращаем внимание читателя на разницу в смысле и обозначениях экспериментальных (выборочных) и теоретических условных средних соответственно $\bar{y}(x)$ и $y_{\text{ср}}(x)$. Строго говоря, на практике теоретических средних мы никогда знать не можем, однако мы опираемся в своем исследовании на тот факт, что в соответствии с законом

И наконец, следует уточнить, в соответствии с каким именно критерием качества аппроксимации неизвестных величин среднедушевых семейных денежных сбережений $y(x)$ и $y_{\text{ср}}(x)$ с помощью функции $\theta_0 + \theta_1 x$ мы будем определять наилучший способ прогноза $y_{\text{ср}}(x)$ по x . Наиболее обоснованное и точное решение этого вопроса опирается на знание вероятностной природы (а именно типа закона распределения вероятностей) остатков ε в модели (10.3). Так, например, известно (см. гл. 2 тома 2), что если предположить, что при любых значениях x распределение вероятностей остатков ε описывается $(0, \sigma^2)$ -нормальным законом и что остатки $\varepsilon(x_i)$, $i = 1, 2, \dots, n$, характеризующие различные наблюдения, статистически независимы, то наименьшая ошибка прогноза $y_{\text{ср}}(x)$ с помощью модели $f(x) \in F$ (т. е. функция $f(x)$ подбирается из класса F) обеспечивается требованием метода наименьших квадратов, в соответствии с которым оценки $\hat{\theta}_0$ и $\hat{\theta}_1$ параметров θ_0 и θ_1 определяются из условия минимизации по θ_0 и θ_1 выражения

$$\sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2. \quad (10.7)$$

В нашем примере явно нарушено условие постоянства дисперсии остатков (см. табл. 10.1), т. е. условная дисперсия $D(\varepsilon \mid \xi = x) = D(\eta - \theta_0 - \theta_1 \xi \mid \xi = x) = \sigma^2(x)$ существенно зависит от значения x . Можно устранить это нарушение, поделив все анализируемые величины, откладываемые по оси η , а следовательно и остатки $\varepsilon(x)$, на значения $\bar{s}(x)$ (являющиеся несмещенными статистическими оценками для $\sigma(x)$), т. е. перейдя к анализу остатков $\tilde{\varepsilon}(x) = \varepsilon(x)/\bar{s}(x)$. Тогда можно показать (с помощью методов, описанных в п. 8.6.1), что гипотеза о $(0; \sigma^2)$ -нормальном характере распределения остатков $\tilde{\varepsilon}(x)$ не противоречит имеющимся в нашем распоряжении данным (представленным в табл. 10.1) и, следовательно, требование (10.7) приводит к необходимости решения экстремальной задачи вида

$$\Delta_n(\theta_0, \theta_1) = \sum_{i=1}^n \left(\frac{y_i - \theta_0 - \theta_1 x_i}{s(x_i)} \right)^2 \rightarrow \min_{\theta_0, \theta_1}, \quad (10.7')$$

т. е. к системе из двух линейных уравнений с двумя неизвестными (θ_0 и

больших чисел (см. п. 6.2) $\bar{y}(x) \rightarrow y_{\text{ср}}(x)$ (по вероятности), когда число наблюдений, по которым подсчитано $\bar{y}(x)$, стремится к бесконечности.

θ_1):

$$\begin{aligned}\frac{\partial \Delta_n(\theta_0, \theta_1)}{\partial \theta_0} &= -2 \sum_{i=1}^n \bar{s}^{-2}(x_i) \cdot (y_i - \theta_0 - \theta_1 x_i) = 0; \\ \frac{\partial \Delta_n(\theta_0, \theta_1)}{\partial \theta_1} &= -2 \sum_{i=1}^n \bar{s}^{-2}(x_i) \cdot x_i \cdot (y_i - \theta_0 - \theta_1 x_i) = 0.\end{aligned}\quad (10.7'')$$

Решение системы (10.7'') дает нам в качестве оценок $\hat{\theta}_0$ и $\hat{\theta}_1$ для неизвестных параметров соответственно θ_0 и θ_1 выражения:

$$\begin{aligned}\hat{\theta}_1 &= \frac{\left(\sum_{i=1}^n \lambda_i\right) \left(\sum_{i=1}^n \lambda_i x_i y_i\right) - \left(\sum_{i=1}^n \lambda_i x_i\right) \left(\sum_{i=1}^n \lambda_i y_i\right)}{\left(\sum_{i=1}^n \lambda_i\right) \left(\sum_{i=1}^n \lambda_i x_i^2\right) - \left(\sum_{i=1}^n \lambda_i x_i\right)^2}, \\ \hat{\theta}_0 &= \frac{\sum_{i=1}^n \lambda_i y_i}{\sum_{i=1}^n \lambda_i} - \hat{\theta}_1 \frac{\sum_{i=1}^n \lambda_i x_i}{\sum_{i=1}^n \lambda_i},\end{aligned}\quad (10.7''')$$

где $\lambda_i = \bar{s}^{-2}(x_i)$.

Расчет по этим формулам с использованием данных табл. 10.1 дает нам ответ на сформулированный выше вопрос 4:

$$\begin{aligned}\hat{\theta}_1 &= 0,685; \\ \hat{\theta}_0 &= -40,360,\end{aligned}$$

так что статистическая оценка искомой зависимости средней величины среднедушевых семейных сбережений $y_{\text{ср}}(x)$ от значения среднедушевого дохода семей доходной группы x имеет в этом случае вид

$$\hat{y}_{\text{ср}}(x) = -40,36 + 0,685 \cdot x.$$

При другой статистической природе остатков ϵ или при отсутствии достаточной информации о типе их вероятностного распределения возможен иной, чем по (10.7), выбор критерия качества аппроксимации Δ_n . Отметим, однако, что наиболее широкое распространение в статистической практике получили именно различные варианты критерия наименьших квадратов (10.7) (см. гл. 2 в томе 2).

Заканчивая обсуждение примера 10.1 и возвращаясь к общему описанию задач статистического исследования зависимостей, отметим, что

функции $f(X) = E(\eta | \xi = X)$, описывающие поведение условных средних результирующего показателя η (вычисленных при значениях объясняющих переменных ξ , зафиксированных на уровне $\xi = X$) в зависимости от изменения X , принято называть *функциями регрессии*.

10.2. Какова конечная прикладная цель статистического исследования зависимостей?

С этого вопроса должно начинаться любое статистическое исследование зависимостей. Ведь от ответа на этот вопрос существенно зависят план исследования, выбор общей структуры математической модели, интерпретация получаемых статистических характеристик и выводов и т. д.

Итак, для чего же строятся математические модели типа (10.3), описывающие статистические зависимости между исследуемыми переменными: результирующими показателями $Y = (y^{(1)}, y^{(2)}, \dots, y^{(m)})$, с одной стороны, и соответствующими объясняющими (экзогенными) переменными $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})$, с другой стороны?

Выделим три основных типа конечных прикладных целей подобных исследований, расположив их как бы по нарастанию глубины проникновения в содержательную сущность анализируемой конкретной задачи.

Тип 1: *Установление самого факта наличия (или отсутствия) статистически значимой связи между Y и X .* При такой постановке задачи статистический вывод имеет двоичную (альтернативную) природу — «связь есть» или «связи нет» — и сопровождается обычно лишь численной характеристикой (измерителем) степени тесноты исследуемой зависимости. Выбор *формы связи* (т.е. класса допустимых решений F и конкретного вида функции $f(X)$ в модели (10.3)) и состава объясняющих переменных X играет подчиненную роль и нацелен исключительно на максимизацию величины этого измерителя степени тесноты связи: исследователю часто не приходится даже «добираться» до конкретного вида функции $f(X)$ и тем более он не претендует на анализ причинных влияний переменных X на результирующие показатели. Этой проблеме посвящена гл. 11.

Тип 2: *прогноз (восстановление) неизвестных значений интересующих нас индивидуальных ($Y(X) = (\eta | \xi = X)$) или средних ($Y_{cp}(X) = E(\eta | \xi = X)$) значений исследуемых результирующих показателей по заданным значениям X соответствующих объясняющих переменных.* При такой постановке задачи статистический вывод включает в себя описание интервала (области) вероятных значений прогнозируемого показателя

теля $Y_{\text{ср}}(X)$ или $Y(X)$ и сопровождается величиной доверительной вероятности P , с которой гарантируется справедливость нашего прогноза, формализуемого с помощью утверждения вида $\{Y(X) \in \Delta[Y(X)]_P\}$ или $\{Y_{\text{ср}}(X) \in \Delta[Y_{\text{ср}}(X)]_P\}$ (см. определение доверительного интервала в п. 7.4 и п. 7.5.4). Как и в предыдущем случае, выбор формы связи (т. е. класса допустимых решений F и конкретного вида функции $f(X)$ в модели (10.3)) и состава предикторов объясняющих переменных X играет подчиненную роль и нацелен исключительно на минимизацию ошибки получаемого прогноза. Однако в данном случае (в отличие от предыдущего) исследователь *существенно использует* значения функции $f(X)$, которые являются отправной точкой при построении прогнозных доверительных интервалов. Последние обычно определяются в форме множества всех тех значений Y , которые удовлетворяют неравенствам

$$f(X) - \varepsilon_P(X, n) \leq Y \leq f(X) + \varepsilon_P(X, n), \quad (10.8)$$

где $\varepsilon_P(X, n)$ — гарантируемая (с вероятностью, не меньшей заданного значения P) максимальная величина ошибки прогноза¹. Таким образом, исследователя интересуют в данном случае *лишь значения* функции $f(X)$, но не ее структура, определяющая, в частности, соотношение удельных весов влияния объясняющих переменных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ на каждый из результирующих показателей $y^{(k)}$ ($k = 1, 2, \dots, m$). Так, например, если при статистическом оценивании неизвестной истинной зависимости

$$f(X) = y_{\text{ср}}(x^{(1)}, x^{(2)}) = 1 + 3x^{(1)} + 5x^{(2)} \quad (10.9)$$

исследователю удалось получить оценку функции $f(X)$ в виде

$$\hat{f}(X) = 1 + 6x^{(1)} - x^{(2)} \quad (10.9')$$

и при этом было установлено, что объясняющие переменные $x^{(1)}$ и $x^{(2)}$ связаны между собой «почти функциональной» линейной зависимостью²

$$x^{(1)} \approx 2x^{(2)}, \quad (10.10)$$

¹ Напоминаем читателю, что f, ε и Y являются m -мерными векторами (см. (10.2)), так что запись (10.8) означает справедливость m соответствующих покомпонентных неравенств.

² Говоря о «почти функциональной» линейной зависимости между $x^{(1)}$ и $x^{(2)}$, мы имеем в виду близость к единице (по абсолютной величине) коэффициента корреляции между этими переменными (см. п. 2.6.6, соотношение (2.33)).

то функция $\hat{f}(X)$ будет обладать хорошими прогностическими свойствами, несмотря на существенное отличие ее коэффициентов при $x^{(1)}$ и $x^{(2)}$ от соответствующих коэффициентов истинной функции $f(X)$. (Обращаем внимание читателя на тот факт, что коэффициенты при $x^{(2)}$ в функциях $f(X)$ и $\hat{f}(X)$ отличаются даже по знаку!) При подстановке заданных значений объясняющих переменных $x^{(1)}$ и $x^{(2)}$ в правые части (10.9) и (10.9'), при условии, что эти значения связаны приближенным соотношением (10.10), мы будем получать совпадающие (или приближенно совпадающие) результаты $f(X)$ и $\hat{f}(X)$, характеризующие усредненную величину $y_{\text{ср}}(X)$ исследуемого результирующего показателя.

Тип 3: *выявление причинных связей между объясняющими переменными X и результирующими показателями Y , частичное управление значениями Y путем регулирования величин объясняющих переменных X .* Такая постановка задачи претендует на проникновение в «физический механизм» изучаемых статистических связей, т. е. в тот самый механизм преобразования «входных» переменных X и ϵ в результирующие показатели Y (см. рис. 10.1), который в большинстве случаев исследователь, не будучи в состоянии его конструктивно описать, вынужден именовать (следуя сложившейся кибернетической терминологии) «черным ящиком».

И при выявлении причинных связей, и при намерении исследователя использовать модели типа (10.3) или (10.4) для управления значениями результирующих показателей $Y_{\text{ср}}(X)$ или $Y(X)$ путем регулирования величин объясняющих переменных X на первый план выходит *задача правильного определения структуры модели* (т. е. выбора общего вида функции $f(X)$), решение которой обеспечивает возможность количественного измерения эффекта воздействия на $Y(X)$ каждой из объясняющих переменных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ *в отдельности*. Однако как раз это место (правильный выбор общего вида функции $f(X)$) и является самым слабым во всей технике статистического исследования зависимостей: к сожалению, не существует стандартных приемов и методов, которые образовывали бы строгую теоретическую базу для решения этой важнейшей задачи (некоторые рекомендации по проведению этого этапа исследования содержатся в п. 10.7).

Заметим, что исследователи, пожалуй, чаще других ставят перед собой именно цели типа 3. И в таких прикладных задачах, как *управление качеством продукции* с помощью регулирования хода технологических процессов, *прогноз и анализ объемов произведенной продукции* по затратам на трудовые ресурсы и капитальные вложения, *построение интегральных целевых функций*, описывающих эффективность функционирования экономических единиц (предприятий, семей) по набору частных

характеристик и др., это вполне оправдано. Однако, к сожалению, далеко не всегда целевые установки исследователей подкреплены объективными возможностями их реализации.

10.3. Математический инструментарий

Методы статистического исследования зависимостей составляют содержание специального раздела *многомерного статистического анализа*, который в свою очередь можно определить как важнейшую составную часть прикладной статистики, содержащую инструментарий для построения оптимальных планов сбора, систематизации и анализа многомерных статистических данных типа (10.1), нацеленный в первую очередь на выявление характера и структуры взаимосвязей между компонентами исследуемого многомерного признака (X, Y) и предназначенный для получения научных и практических выводов. При этом среди $p + m$ компонент исследуемого многомерного признака (X, Y) могут быть: *количественные*, т. е. скалярно измеряющие в определенной шкале степень проявления изучаемого свойства объекта (денежный доход и сбережения семьи, объем валовой продукции, численность работников на предприятии и т. п.); *порядковые* (или *ординальные*), т. е. позволяющие упорядочивать анализируемые объекты по степени проявления в них изучаемого свойства (уровень жилищных условий семьи, квалификационный разряд рабочего, уровень образования работника и т. п.); *классификационные* (или *номинальные*), т. е. позволяющие разбивать обследованную совокупность объектов на не поддающиеся упорядочиванию однородные (по анализируемому свойству) классы (профессия работника, мотив миграции семьи, отрасль промышленности и т. п.). Подразделы многомерного статистического анализа, составляющие математический аппарат статистического исследования зависимостей, формировались и развивались с учетом специфики анализируемых моделей, обусловленной природой изучаемых переменных. Соответствующая специализация этих разделов отражена в табл. 10.2. В ней же указаны главы данной книги, посвященные описанию указанных подразделов.

Из табл. 10.2 видно, что данная книга не охватывает методов исследования зависимостей неколичественного или смешанного (разнотипного) результирующего показателя от количественных или смешанных объясняющих переменных: объемность и специфичность указанной темы обуславливают целесообразность посвящения ей специального издания.

Таблица 10.2

№ п/п	Природа резуль- тирующих пока- зателей (эндоген- ных переменных)	Природа объ- ясняющих (экзо- генных) пере- менных	Название обслу- живающих подраз- делов многомерно- го статистичес- кого анализа	Главы книги, посвя- щенные данном разделам
1	2	3	4	5
1	Количественная	Количественная	Регрессионный и корреляционный анализ	10, 11 (том 1); 1,2,4 (том 2)
2	Количественная	Единственная количественная переменная, ин- терпретируемая как «время»	Анализ времен- ных рядов	3 (том 2)
3	Количественная	Неколичественная (ординальные или номинальные пе- ременные)	Дисперсионный анализ	—
4	Количественная	Смешанная (коли- чественные и неколичествен- ные переменные)	Ковариационный анализ, модели типологической регрессии	—
5	Неколичественная (порядковые, или ординальные, переменные)	Неколичественная (ординальные и номинальные переменные)	Анализ ранговых корреляций и таблиц сопряжен- ности	11 (том 1)
6	Неколичественная (классификацион- ные, или номи- нальные, пере- менные)	Количественная	Дискриминантный анализ, кластер- анализ, таксоно- мия, расщепление смесей распреде- лений	12 (том 1)
7	Смешанная (ко- личественные и неколичественные переменные)	Смешанная (ко- личественные и неколичественные переменные)	Аппарат логи- ческих решаю- щих функций	—

Кроме того, принцип систематизации различных схем, принятый в табл. 10.2, не приспособлен для выделения одного важного (особенно в области социально-экономических приложений) случая, когда связи между количественными переменными X и Y описываются *системой одновременных уравнений*, в которых одни и те же переменные могут играть одновременно (в различных уравнениях системы) и роль результирующих, и роль объясняющих. Этому посвящена теория одновременных эконометрических уравнений, основные результаты которой представлены в гл. 1 и 4 тома 2.

10.4. Некоторые типовые задачи практики эконометрического моделирования

Накопленный опыт практического использования аппарата статистического исследования зависимостей позволяет выделить те типы основных прикладных направлений исследований, в которых этот аппарат работает особенно часто и плодотворно. Если попытаться расщепить общую проблему оптимального управления сложной системой (т. е. центральную проблему кибернетики) на основные составляющие (рис. 10.3), то в качестве этих составляющих как раз и фигурируют именно те направления прикладных исследований, в разработке которых существенную роль играет математический аппарат статистического исследования зависимостей.

Остановимся кратко на роли методов статистического исследования зависимостей в разработке каждого из упомянутых направлений.

I. Нормирование

Общая схема формирования нормативов с использованием методов статистического исследования зависимостей может быть представлена следующим образом. Нормативный показатель играет в моделях типа (10.3)–(10.4) роль результирующей (объясняемой) переменной y , а факторы, участвующие в расчете нормативного показателя, — роль объясняющих переменных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$. Предполагается, что привлечение для расчета норматива y полной системы определяющих его факторов, т. е. такой системы, с помощью которой возможно детерминированное (однозначное) определение величины y , либо принципиально невозможно, либо нецелесообразно из-за чрезмерного усложнения расчетных формул. Поэтому анализируется связь между y и $(x^{(1)}, x^{(2)}, \dots, x^{(p)})$ вида

$$y = f(x^{(1)}, x^{(2)}, \dots, x^{(p)}; \Theta) + \varepsilon, \quad (10.11)$$

где ε — остаточная случайная компонента, обуславливающая возможную погрешность в определении норматива y по известным значениям факторов $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^T$, а $f(X; \Theta)$ — функция из некоторого известного параметрического семейства $F = \{f(X; \Theta)\}$, $\Theta \in A$, однако численное



Рис. 10.3. Основные направления практического использования аппарата статистического исследования зависимостей и центральная проблема кибернетики

значение входящего в ее уравнение параметра Θ (вообще говоря, векторного) неизвестно. С целью подбора «подходящего» значения Θ проводится контрольный эксперимент (наблюдение), в результате которого исследователь получает исходные статистические данные вида (10.1). Далее на основании этих данных проводится необходимый статистический анализ модели (10.11) с целью получения оценки $\hat{\Theta}$ неизвестного параметра Θ и анализа точности полученной расчетной формулы $\hat{Y}_{\text{ср}}(X) = f(X; \hat{\Theta})$, в которой величина условной (экспериментальной) средней $\hat{Y}_{\text{ср}}(X)$ интерпретируется как средний нормативный показатель при значениях определяющих факторов, равных X .

Данный подход использовался, в частности, при разработке методик численности служащих (по различным их функциям) на промышленном предприятии отрасли по набору технико-экономических показателей, характеризующих предприятие, при построении автоматизированных систем нормирования ремонтных работ и в других областях.

II. Прогноз, планирование, диагностика

Отправляясь от общей формулировки задачи статистического исследования зависимостей (см. п. 10.1) и от ее модельной записи (10.11), определим в качестве результирующей переменной y интересующий нас прогнозируемый (планируемый, диагностируемый) показатель, а в качестве объясняющих переменных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ — сопутствующие факторы, значения которых содержат основную информацию о величине этого показателя¹. Наличие остаточной случайной компоненты ϵ , как и прежде, отражает тот факт, что переменные $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ содержат не всю информацию об y , и обуславливает неизбежность погрешности в определении прогнозируемого (планируемого, диагностируемого) показателя по известным значениям объясняющих факторов $x^{(1)}, x^{(2)}, \dots, x^{(p)}$. Исходные статистические данные вида (10.1) исследователь получает, регистрируя одновременно значения y и $(x^{(1)}, \dots, x^{(p)})$ на анализируемых объектах в прошлом (в базовом периоде) или на других объектах, но однородных с анализируемыми.

III. Оценка труднодоступных для непосредственного наблюдения и измерения параметров системы

Восстановление возраста археологической находки по ряду косвенных признаков; *прочности бетона* с помощью косвенных (неразрушающих) методов контроля (например, по отношению диаметров отпечатков на поверхности испытуемого образца бетона и на воздействующем на него эталонном молотке); *денежных сбережений семьи* по ее доходу (в среднедушевом исчислении) — во всех этих ситуациях исследователь вынужден иметь дело с показателями, труднодоступными для непосредственного измерения (они выделены в тексте курсивом). Очевидно, для того чтобы иметь принципиальную возможность статистически выявить связь, су-

¹ В моделях прогноза и планирования в качестве одного из объясняющих факторов $x^{(k)}$ вводится в явном виде «длина прогноза», или «горизонт планирования», t (в единицах времени).

существующую между труднодоступным показателем y и косвенно связанными с ним, но легко поддающимися наблюдению и измерению признаками $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, исследователю необходимо располагать исходными статистическими данными вида (10.1), которые получают с помощью специально организованного контрольного эксперимента или наблюдения. После того как эта связь выявлена (и оценена степень ее точности), она используется для косвенного определения значений труднодоступных показателей лишь по значениям объясняющих переменных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$.

IV. Оценка эффективности функционирования (или качества) анализируемой системы

Пытаясь оценить (в целом) эффективность деятельности отдельного специалиста, подразделения или предприятия, проранжировать страны по некоторому интегральному качеству (например, по качеству жизни населения или по так называемому общему индексу человеческого развития), наконец, проставить балльные оценки спортсмену — участнику командных соревнований в игровых видах спорта за качество его игры в определенном цикле, мы каждый раз по существу решаем (на интуитивном уровне) одну и ту же задачу: отправляясь в своем анализе от набора частных показателей $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, каждый из которых может быть измерен и характеризует какую-нибудь одну частную сторону понятия «эффективность», мы их как бы взвешиваем (т. е. внутренне оцениваем удельный вес их влияния на общее, агрегированное, понятие эффективности) и выходим на некоторый скалярный агрегированный показатель эффективности y . Этот показатель — латентный (скрытый), так как он принципиально не поддается непосредственному измерению (не существует или нам неизвестна объективная шкала, в которой он мог бы быть измерен). Но он с некоторой точностью восстанавливается по значениям частных показателей эффективности $x^{(1)}, x^{(2)}, \dots, x^{(p)}$. Это значит, что между латентным агрегированным показателем y и набором частных критериев эффективности $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ существует статистическая связь типа (10.11).

Главная особенность (и трудность) описываемой ситуации заключается в том, что при получении (сборе) исходной статистической информации вида (10.1) значения результирующего показателя y могут быть получены только с помощью специально организованного экспертного опроса (значения частных критериев эффективности $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, как правило, поддаются непосредственному измерению). Форма экспертной информации о значениях y может быть различной (балльные оценки, упорядочения, парные сравнения). Но только располагая наряду со статистической

информацией об $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^T$ одной из форм соответствующей экспертной информации об y , мы можем статистически построить некоторую аппроксимацию $\hat{y}_{\text{ср}}(X) = f(X; \hat{\Theta})$ для агрегированного критерия эффективности функционирования системы и использовать ее затем в качестве формализованного метода оценки интегрального понятия эффективности (т.е. уже без привлечения экспертов, а лишь по частным критериям $x^{(1)}, x^{(2)}, \dots, x^{(p)}$). Такая модифицированная форма использования аппарата статистического исследования зависимостей предложена в [Айвазян С. А., 1974] и носит название *экспертно-статистического метода построения неизвестной целевой функции* (см. п. 13.5).

В описанную схему вкладывается широкий класс задач теории и практики измерения комплексного понятия «качество» сложной системы: в этих задачах y интерпретируется как агрегированный (комплексный) показатель качества системы, а $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ — как отдельные частные характеристики его качества (надежность, экономичность, удобство пользования, эстетический вид и т. п.). В качестве параметрических семейств $F = \{f(X; \Theta)\}$, привлекаемых при статистическом анализе задач данного типа, чаще других используются функции *линейные*

$$f(X; \Theta) = \theta_0 + \theta_1 x^{(1)} + \dots + \theta_p x^{(p)} \quad (10.12)$$

и степенные

$$f(X; \Theta) = \theta_0 (x^{(1)})^{\theta_1} (x^{(2)})^{\theta_2} \dots (x^{(p)})^{\theta_p}. \quad (10.13)$$

V. Оптимальное регулирование параметров функционирования анализируемой системы, ситуационный анализ

Рассмотрим пример (заимствован из [Айвазян С. А., 1968]). При анализе производительности мартеновских печей на одном из заводов исследовалась, в частности, зависимость между производительностью в тонно/часах (для исключения влияния задержек и простоев часовая производительность мартеновской печи определялась как частное от деления массы плавки на продолжительность периода от начала завадки до выпуска) и процентным содержанием углерода в металле по расплавлению ванны (пробу брали через час после первого скачивания шлака). Результаты замеров по 130 плавкам (т.е. объем n обрабатываемой статистической выборки вида (10.1) равен 130) приведены на рис. 10.4. Очевидно, величины производительности (y_i) и процентного содержания углерода (x_i) подвержены некоторому неконтролируемому разбросу, обусловленному влиянием множества не поддающихся строгому учету и контролю факторов.

Другими словами, последовательность пар чисел $(x_i, y_i), i = 1, 2, \dots, 130$, представляет в данном случае результаты 130 независимых наблюдений двумерной случайной величины (ξ, η) . Однако сквозь кажущуюся хаотичность расположения точек (x_i, y_i) на рис. 10.4 просматривается вполне определенная закономерность зависимости условного среднего значения производительности $y_{\text{ср}}(x) = E(\eta | \xi = x)$ от величины процентного содержания углерода x . Поэтому, располагая статистической зависимостью $y_{\text{ср}}(x)$, мы можем дать рекомендации технологу по оптимальному (с точки зрения максимизации производительности) управлению процессом выплавки: поддерживать процентное содержание углерода в пределах 0,6 – 1,0%.

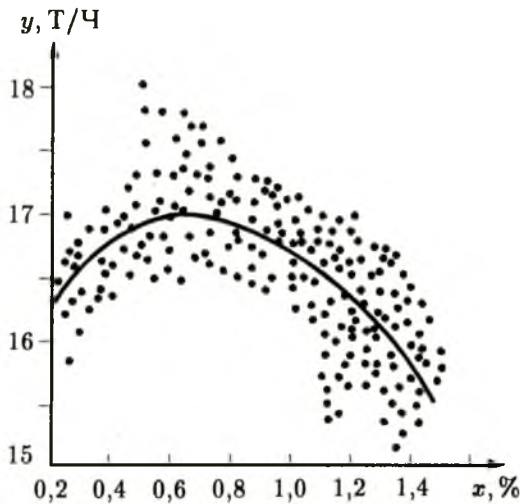


Рис. 10.4. Зависимость производительности (y , т/ч) от процентного содержания углерода (x , %) в металле до расплавления

Мы не случайно начали с этого примера. Использование методов статистического исследования зависимостей в задачах оптимального регулирования хода технологического процесса и построения соответствующих автоматизированных систем управления технологическими процессами можно отнести к примерам грамотных и относительно распространенных актуальных приложений этого аппарата. Общая схема таких приложений предусматривает (в дополнение к приведенному выше частному примеру): а) одновременное рассмотрение нескольких результирующих показателей $y^{(1)}, y^{(2)}, \dots, y^{(p)}$ (производительность, качество продукции, расход сырья и энергии и т. п.) и многих регулируемых параметров тех-

нологического процесса $x^{(1)}, x^{(2)}, \dots, x^{(p)}$; б) возможность сбора исходной статистической информации вида (10.1).

Менее освоенным (но не менее правомерным и актуальным) является этот подход в задачах оптимального регулирования:

- характеристик социально-экономического поведения людей и целых коллективов в ситуациях, когда существует принципиальная возможность выявления статистических связей между этими характеристиками и набором объясняющих (и хотя бы частично регулируемых) факторов;
- структуры и объемов нагрузок и видов заданий в процессе профессиональной подготовки специалистов.

Особо выделим *макро-уровень* подобного моделирования с целью оптимального регулирования параметров функционирования анализируемой системы. В этом случае речь идет об оптимальном регулировании тех макро-параметров национальной экономики, которые поддаются хотя бы частичному управлению и планированию (институциональные и структурные преобразования, налоговая и социальная политика, инвестиционная активность государства и т. п.). Построив и оценив статистические связи, существующие между этими параметрами (так называемыми *экзогенными* переменными), с одной стороны, и результирующими (*эндогенными*, т. е. формирующимися внутри и в ходе функционирования национальной экономики) переменными — с другой, исследователь может, придавая различные значения управляемым параметрам, отслеживать соответствующие реакции на это эндогенных переменных. Т. е. *происходит как бы многократная модельная «прогонка» различных сценариев социально-экономического развития* (такой способ исследования называют также **ситуационным анализом**). Реализуется этот подход, как правило, с помощью *систем регрессионных уравнений* типа (10.3)–(10.4), которые принято называть в эконометрике **системами одновременных уравнений** (см. гл. 1 и 4 тома 2).

10.5. Основные типы зависимостей между количественными переменными

При изучении взаимосвязей между анализируемыми количественными показателями следует установить, к какому именно типу зависимостей относится исследуемая схема. Под типом зависимости мы подразумеваем в данном случае не аналитический вид функции $Y_{cp}(X) = f(X; \Theta)$ в моделях вида (10.11) (о выборе общего аналитического вида функции $f(X; \Theta)$ см. п. 10.7), а природу анализируемых переменных (X, y) и соответственно интерпретацию функции $f(X; \Theta)$ в каждом конкретном случае.

Зависимость между неслучайными переменными (схема А). В этом случае результирующий показатель y детерминированно (т. е. вполне определенно, однозначно) восстанавливается по значениям неслучайных объясняющих переменных $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^T$, т. е. значения y зависят только от соответствующих значений X и полностью ими определяются. Это — обычная схема *чисто функциональной зависимости* между неслучайными переменными, когда y является некоторой функцией от p переменных X (т. е. $y = f(X)$), что является вырожденным случаем зависимостей вида (10.11), когда остаточная случайная компонента ε равна нулю (с вероятностью единица).

Известно, например, что возраст дерева y (в годах) можно однозначно восстановить по числу колец x на срезе его ствола, а именно $y = x$. Примеры адекватного описания реальных зависимостей с помощью чисто функциональных (нестохастических) связей, к сожалению, крайне редки в практике исследований. Кроме того, при проведении их анализа нет необходимости использовать методы вероятностно-статистической теории. Поэтому в дальнейшем изложении мы не будем больше возвращаться к этому типу зависимостей.

Регрессионная зависимость случайного результирующего показателя η от неслучайных объясняющих переменных X (схема В). Природа такой связи может носить двойственный характер: а) регистрация результирующего показателя η неизбежно связана с некоторыми случайными ошибками измерения ε , в то время как предикторные (объясняющие) переменные $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^T$ измеряются без ошибок (или величины этих ошибок пренебрежимо малы по сравнению с соответствующими ошибками измерения результирующего показателя); б) значения результирующего показателя η зависят не только от соответствующих значений X , но и еще от ряда неконтролируемых факторов, поэтому при каждом фиксированном значении X^* соответствующие значения результирующего показателя $\eta(X^*) = (\eta | X = X^*)$ неизбежно подвержены некоторому случайному разбросу.

В этом случае объясняющие переменные X играют роль неслучайного (векторного при $p > 1$) параметра, от которого зависит закон распределения вероятностей (в частности, среднее значение и дисперсия) исследуемого результирующего показателя η . Удобной математической моделью такого рода зависимостей является разложение вида

$$\eta(X) = f(X) + \varepsilon(X). \quad (10.14)$$

Модель (10.14) строится таким образом, что математическое ожидание случайного остатка $\varepsilon(X)$ равно нулю ($E\varepsilon(X) \equiv 0$) тождественно по X ; по-

этому функция $f(X)$ описывает поведение условного среднего $y_{\text{ср}}(X) = E(\eta | X) = f(X)$ в зависимости от X . Предполагается обычно, что при всех X существует конечная дисперсия $\varepsilon(X)$ (т.е. $DE(X) < \infty$), причем величина этой дисперсии, вообще говоря, может зависеть от X (т.е. $DE(X) = \sigma^2(X)$). Подчеркнем то обстоятельство, что в описанной модели (10.14) ни природа случайной компоненты $\varepsilon(X)$, ни соответственно характеристики ее вероятностного распределения никак не связаны со структурой функции $f(X)$ и, в частности, не зависят от значений ее параметра Θ в параметрической записи модели (т.е. когда вместо всех возможных функций $f(X)$ рассматривают какое-либо параметрическое семейство $f(X; \Theta)$, см., например, (10.12), (10.13)).

Если вернуться к примеру 10.1, то можно убедиться, что он хорошо укладывается в рамки модели (10.14). Для этого следует лишь заметить, что имевшаяся в этом примере возможность контролировать значения объясняющей (предикторной) переменной ξ по существу переводит эту переменную из категории случайных величин в категорию неслучайных (контролируемых) параметров модели. Дальнейший анализ примера 10.1 (см. табл. 10.1, формулу (10.5) и рис. 10.2) подсказал нам следующую конкретизацию допущений о природе составных частей модели (10.14):

$$\begin{aligned} y_{\text{ср}}(x) &= E(\eta | X) = f(x) = \theta_0 + \theta_1 x; \\ \sigma^2(x) &= DE(x) = \sigma_0^2 \cdot (y_{\text{ср}}(x))^2, \end{aligned} \quad (10.15)$$

где σ_0 — константа, не зависящая от x .

Подчеркнем тот факт, что именно зависимость по схеме В, т.е. регрессионная зависимость результирующего показателя от *неслучайных величин* объясняющих переменных, лежит в основе большинства эконометрических моделей и, в частности, используется при построении и анализе классической и обобщенной линейной модели регрессии (см. гл. 2 тома 2).

Корреляционно-регрессионная зависимость между случайными векторами η — результирующим показателем и ξ — объясняющей переменной (схема С). В данном типе моделей и компоненты вектора результирующего показателя η , и компоненты вектора объясняющих переменных ξ зависят от множества неконтролируемых факторов, так что являются случайными по своей физической сущности. Мы уже сталкивались с такой ситуацией в примере, в котором исследовалась связь между производительностью мартеновских печей и процентным содержанием углерода в металле (см. рис. 10.4). Зависимости такого типа вообще характерны для описания хода технологических процессов, реальные зна-

чения параметров которых $\xi = (\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(p)})^\top$, равно как и характеризующие их результирующие показатели $\eta = (\eta^{(1)}, \eta^{(2)}, \dots, \eta^{(m)})^\top$, как правило, флуктуируют случайным (но взаимосвязанным) образом около установленных номиналов.

В подобных ситуациях оказывается полезным рассмотреть разложение исследуемого результирующего показателя η на две случайные составляющие по формуле типа (10.3). Первая из них определяется некоторой (векторнозначной) функцией f от объясняющей переменной ξ , а вторая отражает остаточные влияния неучтенных случайных факторов на анализируемый результирующий показатель η . Итак,

$$\eta = f(\xi) + \varepsilon. \quad (10.16)$$

В частном случае единственного результирующего показателя ($m = 1$) и линейного вида функции $f(\xi)$ имеем:

$$\eta = \theta_0 + \sum_{k=1}^p \theta_k \xi^{(k)} + \varepsilon. \quad (10.17)$$

Подразумевая, как и прежде, под $y_{cp}(X) = E(\eta | \xi = X)$ условное математическое ожидание результирующего показателя η (при условии, что объясняющая переменная ξ приняла значение, равное X), мы от (10.17) приходим к линейному уравнению регрессии

$$y_{cp}(X) = \theta_0 + \sum_{k=1}^p \theta_k x^{(k)}. \quad (10.18)$$

Возможны случаи, когда вторая (остаточная) компонента в разложении (10.16) с полной мерой достоверности (т.е. с вероятностью единица) равна нулю. При этом исследуемые случайные величины η и ξ оказываются связанными *чисто функциональной зависимостью* $\eta = f(\xi)$, но ее следует отличать от функциональной зависимости неслучайных переменных (см. выше, схема А).

П р и м е р 10.2. Рис. 10.5 иллюстрирует связь между вакуумом в печи для обжига стекла ξ и процентом брака η в стекольном производстве [Айвазян С. А., 1968].

Случайные изменения свойств сырья, а также ряда неконтролируемых факторов приводят к случайным колебаниям обеих исследуемых переменных. Однако расположение точек на рис. 10.5 свидетельствует о том, что эти колебания взаимосвязаны, подчинены вполне определенной закономерности: «облако» рассеяния вытянуто вдоль некоторой прямой, не

параллельной ни одной из координатных осей. Все это подтверждает целесообразность разложения случайной величины η по формуле (10.16) и исследования связи между η и ξ , которая в этом случае носит название корреляционной. К перечисленным вопросам регрессионного анализа (построение конкретного вида зависимости между переменными, различные оценки ее точности) в этом случае присоединяется круг вопросов, связанных с исследованием степени тесноты связи между этими переменными. Совокупность методов, позволяющих решать эти вопросы, принято называть корреляционным анализом (см. гл. 11).

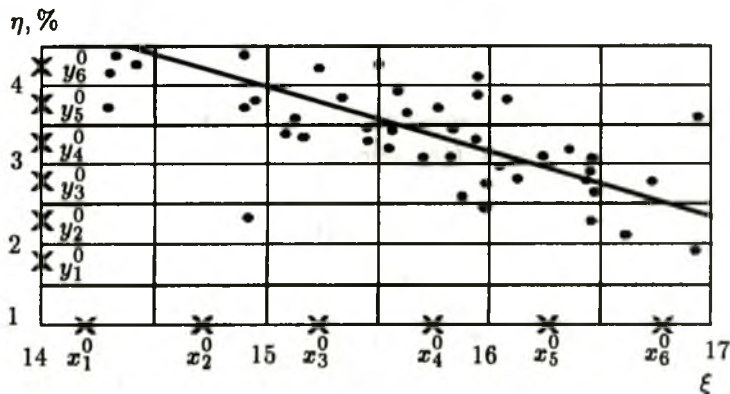


Рис. 10.5. Графическое представление данных по связи вакуума в печи для обжига стекла (ξ) и процента брака в стекольном производстве (η)

Зависимости структурного типа, или зависимости по схеме конфлюэнтного анализа (схема D). В описываемой ниже схеме речь идет о восстановлении искомым зависимостей по искаженным наблюдениям анализируемых переменных, причем в отличие от регрессионной схемы в искаженными оказываются при наблюдении не только значения результирующего показателя, но и значения *объясняющих (предикторных) переменных* $x^{(1)}, x^{(2)}, \dots, x^{(p)}$. Этот тип связей упоминается в специальной литературе как *структурные зависимости* [Кендалл М. Дж., Стьюарт А., 1973, с. 500–557] или как зависимости по схеме *конфлюэнтного анализа* [Айвазян С. А., 1979].

Таким образом, конфлюэнтный анализ предоставляет совокупность методов математико-статистической обработки данных, относящихся к анализу априори постулируемых функциональных связей между количественными (случайными или неслучайными) переменными $Y =$

$(y^{(1)}, \dots, y^{(m)})^\top$ и $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^\top$ в условиях, когда наблюдаются не сами переменные, а случайные величины

$$\begin{aligned} \xi_i^{(k)} &= x_i^{(k)} + \varepsilon_{x_i}^{(k)}, \quad k = 1, 2, \dots, p; \\ \eta_i^{(j)} &= y_i^{(j)} + \varepsilon_{y_i}^{(j)}, \quad j = 1, 2, \dots, m; \quad i = 1, 2, \dots, n, \end{aligned} \quad (10.19)$$

где $\varepsilon_{x_i}^{(k)}$ и $\varepsilon_{y_i}^{(j)}$ — случайные ошибки измерений соответственно переменных $x^{(k)}$ и $y^{(j)}$ в i -м наблюдении, а n — общее число наблюдений. При этом общий вид исследуемых функциональных (структурных) связей

$$\begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{pmatrix} = \begin{pmatrix} f^{(1)}(x^{(1)}, \dots, x^{(p)}; \Theta) \\ \vdots \\ f^{(m)}(x^{(1)}, \dots, x^{(p)}; \Theta) \end{pmatrix} \quad (10.20)$$

между ненаблюдаемыми, а точнее, *наблюдаемыми с ошибками* переменными считается заданным (неизвестным является лишь значение векторного параметра $\Theta = (\theta_1, \dots, \theta_N)$, участвующего в уравнениях искомых зависимостей (10.20)).

10.6. Основные этапы статистического исследования зависимостей

Весь процесс статистического исследования интересующих нас зависимостей удобно разложить на основные этапы. Эти этапы ниже описаны в соответствии с хронологией их реализации, однако некоторые из них находятся, в плане хронологическом, в *соотношении итерационного взаимодействия*: результаты реализации более поздних этапов могут содержать выводы о необходимости повторной «прогонки» (с учетом добытой на предыдущих этапах новой информации) уже пройденных этапов (см., например, схему взаимодействия этапов 3, 4, 5 и 6 на рис. 10.6).

Излагаемая ниже схема приспособлена в основном для исследования зависимостей между количественными переменными, однако с минимальными (и очевидными) модификациями она «работает» и при статистическом анализе связей между неколичественными и разнотипными переменными.

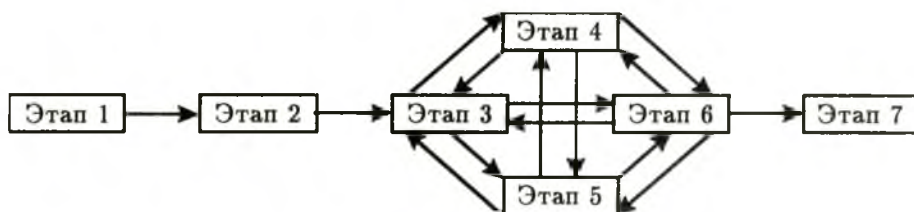


Рис. 10.6. Схема хронологически-итерационных взаимосвязей основных этапов статистического исследования зависимостей

Этап 1 (постановочный). Прежде всего исследователь должен определить:

1) элементарную единицу статистического обследования, или элементарный объект исследования O (это может быть страна, город, отрасль, предприятие, семья, индивидум, пациент, технологический процесс, сложное техническое изделие и т. д.);

2) набор показателей $(x^{(1)}, x^{(2)}, \dots, x^{(p)}; y^{(1)}, \dots, y^{(m)})$, регистрируемых на каждом из статистически обследованных объектов, с подразделением их на «входные» (объясняющие) и «выходные» (результатирующие) и, если это необходимо, с четким определением способа их измерения; таким образом, на этом этапе каждому элементарному объекту исследования ставится в соответствие перечень анализируемых показателей, т. е.

$$O \Leftrightarrow (x^{(1)}, x^{(2)}, \dots, x^{(p)}; y^{(1)}, y^{(2)}, \dots, y^{(m)});$$

3) конечные прикладные цели исследования (см. п. 10.2), тип исследуемых зависимостей (см. п. 10.5) и желательную форму статистических выводов (а иногда и степень их точности);

4) совокупность элементарных объектов исследования, на которую мы хотим распространить справедливость действия выявленных в результате анализа статистических зависимостей (если, например, элементарная единица — семья, то анализируемой совокупностью могут быть семьи определенной социальной группы населения или семьи определенной страны и т. д.);

5) общее время и трудозатраты, отведенные на планируемое исследование и коррелированные с ними временная протяженность и объем необходимого статистического обследования (какую часть анализируемой совокупности подвергнуть статистическому обследованию, производить статистическое обследование в статистическом или динамическом режиме и

т. д.). Заметим, что именно на этом этапе решаются задачи в) и 1, описанные в п. 10.1.

В решении всех перечисленных вопросов первого этапа исследования главную роль, бесспорно, должен играть «заказчик», т. е. специалист той предметной области, для которой планируется проведение этого исследования.

Этап 2 (информационный). Он состоит в проведении сбора необходимой статистической информации вида (10.1). При этом возможны две принципиально различные ситуации: 1) исследователь имеет возможность заранее спланировать выборочное обследование части анализируемой совокупности — выбрать способ отбора элементарных единиц статистического обследования (случайный, пропорциональный, расслоенный и т. д., см. п. 6.1), хотя бы по части объясняющих переменных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ назначить уровни их значений, при которых желательно произвести эксперимент или наблюдения (условия *активного эксперимента*); 2) исследователь получает исходные данные такими, какими они были собраны без его участия (условия *пассивного эксперимента*). В любом случае «на выходе» этого этапа исследователь располагает исходными статистическими данными вида (10.1), т. е. каждому (i -му) из статистически обследованных элементарных объектов O_i поставлен в соответствие конкретный вектор характеризующих его «входных» и «выходных» показателей:

$$O_i \Longleftrightarrow (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)}; y_i^{(1)}, y_i^{(2)}, \dots, y_i^{(m)}), \quad i = 1, 2, \dots, n$$

(здесь n — общее число статистически обследованных элементарных объектов, т. е. объем выборки).

Говоря о проведении сбора статистических данных, мы не включаем сюда разработку методологии и системы показателей отображаемого объекта: эта работа предполагает профессионально-предметное (экономическое, техническое, медицинское и т. д.) изучение сущности решаемых задач статистического исследования зависимостей, поэтому относится к компетенции соответствующей предметной статистики (экономической и т. д.) и входит в задачи 1-го этапа исследований.

Этап 3 (корреляционный анализ). Этот этап нацелен на решение задачи 2 (см. п. 10.1), он позволяет ответить на вопросы, имеется ли вообще какая-либо связь между исследуемыми переменными, какова структура этих связей и как измерить их тесноту? Описанию методов, с помощью которых проводится такой статистический анализ, посвящена гл. 11. Поскольку перечисленные выше вопросы решаются с помощью вычисления и анализа соответствующих корреляционных характеристик, содержание этапа можно определить как проведение корреляционного анализа.

Этап достаточно полно оснащен необходимым математическим аппаратом и программным обеспечением, поэтому может быть почти полностью автоматизирован.

Этап 4 (определение класса допустимых решений). Главной целью исследователя на этом этапе является определение общего вида, структуры искомой связи между Y и X , или, другими словами, описание класса функций F , в рамках которого он будет производить дальнейший поиск конкретного вида интересующей его зависимости (см. задачи а) и 3 в п. 10.1). Чаще всего это описание дается в форме некоторого параметрического семейства функций $f(X; \Theta)$, поэтому и этап этот называют также *этапом параметризации модели*. Так, определив в примере 10.1, что поиск зависимости среднедушевых семейных сбережений $y_{ср}$ от величины их среднедушевого дохода x мы будем производить в классе $F = \{\theta_0 + \theta_1 x\}$ линейных функций, мы тем самым завершили четвертый этап исследования (но конкретных числовых значений параметров θ_0 и θ_1 мы к этому моменту еще не знаем).

Следует отметить, что, являясь узловым, в определенной мере *решающим* звеном во всем процессе статистического исследования зависимостей, этот этап в то же время находится в наименее выгодном положении по сравнению с другими этапами (с позиций наличия строгих и законченных математических рекомендаций по его реализации). Поэтому его реализация требует работы специалиста соответствующей предметной области (экономики, техники, медицины и т. д.) и математика-статистика, направленной на как можно более глубокое проникновение в «физический механизм» исследуемой связи. Подходам и методам проведения этого этапа исследований посвящен следующий пункт этой главы.

Существует подход к исследованию моделей регрессии, не требующий предварительного выбора параметрического семейства функций F , в рамках которого проводится дальнейший анализ. Речь идет о так называемых *непараметрических (или частично-параметрических)* методах исследования регрессионных зависимостей. Однако возникающие при их реализации проблемы (необходимость иметь очень большие объемы исходных статистических данных, выбор сглаживающих функций — «окон» и параметров масштаба, выбор порядка сплайна, числа и положения «узлов» и т. п.) сопоставимы по своей сложности с проблемами, возникающими при реализации этапа 4.

Следующие два этапа — 5-й и 6-й — связаны с проведением определенного объема вычислений и реализуются по существу параллельно.

Этап 5 (анализ мультиколлинеарности предсказывающих переменных и отбор наиболее информативных из них). Под явлением мультикол-

линейности в регрессионном анализе понимается наличие тесных статистических связей между объясняющими переменными $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, что, в частности, проявляется в близости к нулю (слабой обусловленности) определителя их корреляционной матрицы, т.е. матрицы размера $p \times p$, составленной из парных коэффициентов корреляции $r_{ij} = r(x^{(i)}, x^{(j)})$ (см. п. 2.6.6, соотношение (2.32)). Поскольку этот определитель входит в знаменатель для важных характеристик анализируемых моделей (см. гл. 2 тома 2), то мультиколлинеарность создает трудности и неудобства при статистическом исследовании по меньшей мере в двух направлениях:

а) в реализации необходимых вычислительных процедур и, в частности, в крайней неустойчивости получаемых при этом числовых характеристик анализируемых моделей (так, коэффициенты при объясняющих переменных в моделях типа (10.12), (10.13) и др. могут изменяться в несколько раз и даже менять знак при добавлении (или исключении) к массиву исходных статистических данных одного-двух объектов или одной-двух объясняющих переменных);

б) в содержательной интерпретации параметров анализируемой модели, что играет решающую роль в ситуациях, когда конечной целью исследования является цель типа 3 («выявление причинных связей» и т.д., см. п. 10.2, соотношения (10.9) и (10.9')).

Поэтому исследователь старается перейти к такой новой системе объясняющих переменных (отобранных из числа исходных переменных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ или представленных в виде некоторых их комбинаций), в которой эффект мультиколлинеарности уже не имел бы места. Этап проводится в основном силами математиков-статистиков с подключением (в самом его конце) специалистов соответствующей предметной области для выбора из нескольких предложенных вариантов наборов объясняющих переменных одного, наиболее легко и естественно интерпретируемого.

Рекомендации по проведению этого этапа даны в гл. 2 тома 2.

Этап 6 (вычисление оценок неизвестных параметров, входящих в исследуемое уравнение статистической связи). Итак, в результате проведения предыдущих этапов были решены, в частности, следующие задачи:

а) определены результирующие и объясняющие переменные и тип исследуемой зависимости (B, C или D , см. п. 10.5);

б) собрана и подготовлена к счету исходная статистическая информация вида (10.1);

в) изучены характер и теснота статистических (корреляционных) связей между исследуемыми переменными;

г) выбран класс допустимых решений F , т.е. класс (или параметри-

ческое семейство) функций $f(X)$, в рамках которого будет подбираться наилучшая (в определенном смысле) аппроксимация $\hat{f}(X)$ искомой зависимости типа (10.14), (10.16) или (10.20).

Теперь можно приступить к определению этой наилучшей аппроксимации $\hat{f}(X)$, которая является решением оптимизационной задачи вида

$$\hat{f}(X) = \arg \operatorname{extr}_{f \in F} \Delta_n(f), \quad (10.21)$$

где функционал $\Delta_n(f)$ задает критерий качества аппроксимации результирующего показателя η (или Y) с помощью функции $f(X)$ из класса F . Выбор конкретного вида этого функционала опирается на знание вероятностной природы остатков ϵ в моделях типа (10.14) и (10.16), причем он строится, как правило, в виде некоторой функции от невязок $\hat{\epsilon}_1^{(k)}, \hat{\epsilon}_2^{(k)}, \dots, \hat{\epsilon}_n^{(k)}$ ($k = 1, 2, \dots, m$), где $\hat{\epsilon}_i^{(k)} = y_i^{(k)} - f^{(k)}(X_i)$ (один из распространенных вариантов такого функционала, а именно *функционал метода наименьших квадратов*, упоминается в примере 10.1, см. соотношение (10.7')). Если в качестве класса F задаются некоторым параметрическим семейством функций $\{f(X; \Theta)\}$, то задача (10.1) сводится к подбору (статистическому оцениванию) значений параметров $\hat{\Theta}$, на которых достигается экстремум по Θ функционала $\Delta_n(f(X; \Theta))$, а соответствующие модели называют *параметрическими*. Эта часть исследования хорошо оснащена необходимым математическим аппаратом и соответствующим программным обеспечением (см. гл. 2 и 4 тома 2).

Этап 7 (анализ точности полученных уравнений связи). Исследователь должен отдавать себе отчет в том, что найденная им в соответствии с (10.1) аппроксимация $\hat{f}(X)$ неизвестной теоретической функции $f_T(X)$ из соотношений типа (10.14) или (10.16) (называемая эмпирической функцией регрессии, см. гл. 2 тома 2) является лишь некоторым приближением истинной зависимости $f_T(X)$ ¹. При этом погрешность δ в описании неизвестной истинной функции $f_T(X)$ с помощью $\hat{f}(X)$ в общем случае состоит из двух составляющих: а) ошибки аппроксимации δ_F и б) ошибки выборки $\delta(n)$. Величина первой зависит от успеха в реализации этапа 4, т.е. от правильности выбора класса допустимых решений F . В частности, если класс F выбран таким образом, что включает в себя и неизвестную истинную функцию f (т.е. $f_T(X) \in F$), то ошибка аппроксимации $\delta_F = 0$. Но даже в этом случае остается случайная составляющая (ошиб-

¹ В дальнейшем, говоря о вектор-функции $f_T(X)$, вектор-погрешности δ и векторе результирующих показателей $Y(X)$, мы будем иметь в виду каждую из их компонент в отдельности.

ка выборки) $\delta(n)$, обусловленная *ограниченностью* выборочных данных вида (10.1), на основании которых мы подбираем функцию $\hat{f}(X)$ (оцениваем ее параметры). Очевидно, уменьшить ошибку выборки мы можем за счет увеличения объема n обрабатываемых выборочных данных, так как при $f_T(X) \in F$ (т. е. при $\delta_F = 0$) и правильно выбранных методах статистического оценивания (т. е. при правильном выборе оптимизируемого функционала качества модели $\Delta_n(f)$) ошибка выборки $\delta(n) \rightarrow 0$ (по вероятности) при $n \rightarrow \infty$ (свойство состоятельности используемой процедуры статистического оценивания неизвестной функции $f_T(X)$).

Соответственно на данном этапе приходится решать следующие основные задачи анализа точности полученной регрессионной зависимости:

1) в случае $F = \{f(X; \Theta)\}$ и $f_T(X) \in F$, т. е. когда класс допустимых решений задается параметрическим семейством функций и включает в себя неизвестную теоретическую функцию регрессии $f_T(X)$, при заданных доверительной вероятности P и объеме выборки n указать такую предельную (гарантированную) величину погрешности $\delta_{P,n}(\theta_k)$ для любой компоненты неизвестного векторного параметра Θ , что

$$|\theta_k - \hat{\theta}_k| \leq \delta_{P,n}(\theta_k)$$

с вероятностью, не меньшей, чем P (здесь θ_k — истинное значение k -й компоненты неизвестного параметра Θ , а $\hat{\theta}_k$ — его статистическая оценка);

2) при заданных доверительной вероятности P , объеме выборки n и значениях объясняющих переменных X указать такую предельную (гарантированную) величину погрешности $\delta_{P,n}(Y_{cp}(X))$, что

$$|Y_{cp}(X) - \hat{f}(X)| < \delta_{P,n}(Y_{cp}(X))$$

с вероятностью, не меньшей, чем P (здесь $Y_{cp}(X) = E(\eta | X)$ — неизвестное *условное среднее* значение исследуемого результирующего показателя при значениях объясняющих переменных, равных X , а $\hat{f}(X)$ — построенная в соответствии с (10.21) эмпирическая функция регрессии);

3) при заданных доверительной вероятности P , объеме выборки n и значениях объясняющих переменных X указать такую предельную (гарантированную) величину погрешности $\delta_{P,n}(Y(X))$, что

$$|Y(X) - \hat{f}(X)| \leq \delta_{P,n}(Y(X))$$

с вероятностью, не меньшей, чем P (здесь $Y(X)$ — прогнозируемое *индивидуальное* значение исследуемого результирующего показателя при значениях объясняющих переменных, равных X).

Описание методов анализа точности исследуемых регрессионных моделей содержится в гл. 2 тома 2.

Заметим в заключение, что часть исследования, объединяющая этапы 4, 5, 6 и 7, принято называть *регрессионным анализом*.

10.7. Выбор общего вида функции регрессии

Итак, как мы видим (см. предыдущий пункт), собственно регрессионный анализ, т.е. конструирование по исходным данным вида (10.1) неизвестной функции регрессии $f(X) = E(\eta | \xi = X)$, начинается с этапа 4, т.е. с выбора семейства допустимых решений F — класса функций, в рамках которого предполагается вести поиск наилучшей аппроксимации $\hat{f}(X)$ для $f(X)$ ¹.

Наиболее распространенными в статистической практике являются *параметрические регрессионные схемы*, когда в качестве класса допустимых решений выбирается некоторое параметрическое семейство функций

$$F = \{f(X; \Theta)\}_{\Theta \in \Gamma}. \quad (10.22)$$

В этом случае дальнейший поиск аппроксимации $\hat{f}(X)$ сводится к наилучшему (в смысле заданного критерия адекватности, см. выше, этап 6, соотношение (10.21)) подбору неизвестного значения параметра Θ , что в свою очередь осуществляется с помощью *полностью формализованного* алгоритма решения соответствующей оптимизационной задачи, составляющей математическую основу процедуры, называемой *статистическим оцениванием параметра* (см. гл. 7).

Но до перехода к процедуре статистического оценивания неизвестного значения параметра мы должны сделать и обосновать выбор типа параметрического семейства (10.22). Так, например, в качестве класса допустимых решений можно использовать

$$\text{линейные функции: } f(X; \Theta) = \theta_0 + \sum_{k=1}^p \theta_k \cdot x^{(k)}; \quad (10.22')$$

$$\text{степенные функции: } f(X; \Theta) = \theta_0 (x^{(1)})^{\theta_1} (x^{(2)})^{\theta_2} \dots (x^{(p)})^{\theta_p}; \quad (10.22'')$$

¹ Начиная с этого момента мы будем рассматривать случай $m = 1$ общей постановки задачи статистического исследования зависимостей, т.е. случай *единственного* результирующего показателя y .

алгебраические полиномы степени $m \geq 2$:

$$f(X; \Theta) = \theta_0 + \sum_{k=1}^p \theta_k \cdot x^{(k)} + \sum_{k_1=1}^p \sum_{k_2=1}^p \theta_{k_1 k_2} x^{(k_1)} \cdot x^{(k_2)} \\ + \dots + \sum_{k_1=1}^p \dots \sum_{k_m=1}^p \theta_{k_1 k_2 \dots k_m} \cdot x^{(k_1)} \cdot x^{(k_2)} \dots x^{(k_m)} \quad (10.22''')$$

и т. д.

Следует подчеркнуть, что этап 4 (см. п. 10.6), т.е. этап исследования, посвященный выбору общего вида функции регрессии (параметризация модели), бесспорно, является *ключевым*: от того, насколько удачно он будет реализован, решающим образом зависит точность восстановления неизвестной функции регрессии $f(X)$. В то же время приходится признать, что этот этап находится, пожалуй, в самом невыгодном положении: к сожалению, не существует системы стандартных рекомендаций и методов, которые образовывали бы строгую теоретическую базу для его наиболее эффективной реализации.

Остановимся на некоторых рекомендациях, связанных с реализацией трех основных моментов, учет которых необходим при решении проблемы выбора общего вида функции регрессии: 1) максимальное использование априорной информации о *содержательной* (физической, экономической, социологической и т. п.) *сущности* анализируемой зависимости; 2) предварительный анализ *геометрической структуры* исходных данных вида (10.1), на основании которых конструируется искомая зависимость; 3) различные *статистические приемы* обработки исходных данных, позволяющие сделать наилучший выбор из нескольких сравниваемых вариантов.

10.7.1. Использование априорной информации о содержательной сущности анализируемой зависимости

Анализируя содержательную сущность изучаемой зависимости, исследователь еще *до обращения* к исходным статистическим данным может (и должен!) попытаться ответить на ряд вопросов по поводу характера искомой регрессионной связи:

а) будет ли искомая функция $f(X)$ монотонной или она должна иметь один экстремум (может быть, несколько)?

б) следует ли ожидать стремления (в процессе $x^{(k)} \rightarrow \infty$) $f(X)$ к асимптомам (по одной или нескольким предикторным переменным) и какова их содержательная интерпретация? Так, например, если $f(X)$ — сред-

ний объем благ определенного вида, потребляемых семьями группы X по доходам, то, очевидно, при $X \rightarrow \infty$ следует ожидать «насыщения», т. е. $f(X)$ будет стремиться (снизу) к горизонтальной асимптоте (см. п. 4 в табл. 10.3);

в) какова принципиальная природа воздействия объясняющих переменных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ на формирование результирующего показателя y — аддитивная или мультипликативная? Так, например, многие схемы зависимостей в экономике характеризуются мультипликативной природой воздействия предикторов на y (см. п. 1–3 в табл. 10.3);

г) не диктует ли содержательный смысл анализируемой зависимости обязательное прохождение графика искомой функции $f(X)$ через одну или несколько априори заданных точек в исследуемом факторном пространстве (X, y) ?

Поясним необходимость и возможность максимального извлечения информации об общем виде анализируемой функции регрессии $f(X)$ из соображений профессионально-теоретического характера на примере.

Пример 10.3. На рис. 10.7 представлены 63 результата специального эксперимента [Езекиэл М., Фокс К., с. 57]. Расположение точек на рис. 10.7 не дает ответа на вопрос, описывать ли зависимость между скоростью автомобиля (x миль/ч) и расстоянием (y футов), пройденным им после поданного сигнала об остановке, линейной или параболической зависимостью.

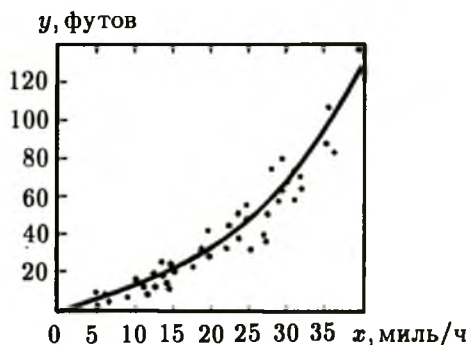


Рис. 10.7. График зависимости тормозного пути автомобиля (y) от скорости его движения (x)

Этот вопрос остается без ответа и после построения соответствующих кривых и применения известных статистических критериев, предназначенных решать, насколько хорошо согласуются кривые с экспериментальными данными. Однако несложные рассуждения профессионально-

теоретического характера все-таки позволяют сделать этот выбор. Действительно, для каждого отдельного автомобиля и водителя расстояние, пройденное до остановки, определяется в основном тремя факторами: скоростью автомобиля (x) в момент подачи сигнала об остановке, временем реакции на этот сигнал водителя (θ_1 , ч) и тормозами автомобиля. Автомобиль успеет пройти путь $\theta_1 x$ до момента включения водителем тормозов и еще $\theta_1 \cdot x^2$ после этого момента, поскольку согласно элементарным физическим законам теоретическое расстояние, пройденное до остановки с момента торможения, пропорционально квадрату скорости.

Итак, $f(x) = \theta_1 x + \theta_2 x^2$, что после оценивания θ_1 и θ_2 с помощью МНК (см. гл. 7) дает $f(x) = 0,76x + 0,056x^2$.

10.7.2. Предварительный анализ геометрической структуры исходных данных

При выяснении вопроса о параметрическом виде исследуемой зависимости, как правило, идут от простого к сложному. Простейшей же аппроксимацией неизвестной функции регрессии $f(X) = E(\eta \mid \xi = X)$ является, естественно, *линейная модель*, т.е. функция вида

$$f_{\Lambda}(X) = \theta_0 + \theta_1 x^{(1)} + \dots + \theta_p x^{(p)}. \quad (10.23)$$

Поэтому при *предварительном* анализе характера исследуемых зависимостей (т.е. проведения вычислительных процедур по оценке неизвестных значений параметров, входящих в гипотетичные уравнения связей) ограничиваются некоторыми приближенными эвристическими приемами, связанными в основном с изучением «геометрии» парных корреляционных полей и визуальной проверкой их линейности.

Содержание геометрического анализа парных корреляционных полей. Под *корреляционным полем* переменных (u, v) понимается графическое представление имеющихся измерений $(u_1, v_1), (u_2, v_2), \dots, (u_n, v_n)$ этих переменных в плоскости (u, v) . Мы уже неоднократно имели дело с корреляционными полями (см. рис. 10.2, 10.4, 10.5, 10.7).

Анализ парных корреляционных полей состоит обычно в следующем:

а) построение на основании имеющихся исходных данных вида (10.1) корреляционных полей для всевозможных пар переменных вида $(x^{(j)}, x^{(k)})$ и $(x^{(j)}, y)$, отобранных из набора всех $p + 1$ исследуемых признаков $(x^{(1)}, x^{(2)}, \dots, x^{(p)}; y)$; всего таких пар будет, очевидно, $p(p + 1)/2$, однако процесс этот легко автоматизируется с помощью современных вычислительных средств;

б) визуальное прослеживание характера вытянутости каждого корреляционного поля: эллипсоидально-линейное (см. рис. 10.6), нелинейно-монотонное (см. рис. 10.7), с наличием одного или нескольких экстремумов (см. рис. 10.4) и т. п.;

в) изучение поведения условных средних значений результирующего показателя при изменении величины переменной, откладываемой по оси абсцисс и играющей роль предикторной (см. рис. 10.2); для этого (если значения предикторной переменной *неконтролируемы* в ходе наблюдения или эксперимента) предварительно разбивают диапазон значений объясняющей переменной на *интервалы группирования* (см. п. 6.1) и подсчитывают средние значения ординат тех точек-наблюдений, которые попали в общий интервал группирования.

В результате такого анализа обычно получают формулировку нескольких рабочих гипотез об общем виде искомой зависимости, окончательная проверка которых и выбор наиболее адекватной из них осуществляются (при отсутствии априорных сведений содержательного характера) с помощью соответствующих *математико-статистических* методов. Описание наиболее эффективных, с нашей точки зрения, приемов такого типа приводятся в п. 10.7.3.

10.7.3. Статистические критерии проверки гипотез об общем виде функции регрессии

Подчеркнем сразу, что описанные ниже критерии проверки справедливости сделанного выбора общего вида искомой функции регрессии не могут ответить на вопрос: является ли проверяемый гипотетичный вид зависимости наилучшим, единственно верным? Они лишь подтверждают факт непротиворечивости проверяемого вида функции регрессии имеющимся у исследователя исходным данным (10.1) либо отвергают обсуждаемую гипотетичную форму зависимости как не соответствующую этим данным.

1. *Общий приближенный критерий, основанный на группированных данных* (или при наличии нескольких наблюдений при каждом фиксированном значении аргумента). Пусть высказана гипотеза об общем виде функции регрессии $H_0: E(\eta | \xi = X) = f_a(X; \theta_1, \theta_2, \dots, \theta_k)$ ($f_a(X; \Theta)$ — известная функция, $(\theta_1, \theta_2, \dots, \theta_k) = \Theta$ — неизвестные числовые параметры) и пусть вычислены (например, с помощью метода наименьших квадратов, см. гл. 2 тома 2) оценки $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ неизвестных параметров, входящих в описание уравнения регрессии. При группировке данных (или при проведении эксперимента) мы должны соблюдать требование, в соот-

ветствии с которым число интервалов группирования (или число различных значений аргумента, в которых производились наблюдения) s должно обязательно превосходить число неизвестных параметров k , т. е. $s - k \geq 1$.

Если высказанная гипотеза об общем виде зависимости является правильной, то статистика

$$v^2 = \frac{\frac{1}{s-k} \sum_{i=1}^s \nu_i |\bar{y}_i - f_a(X_i^0; \hat{\Theta})|^2}{\frac{1}{n-s} \sum_{i=1}^s \sum_{j=1}^{\nu_i} |y_{ij} - \bar{y}_i|^2} \quad (10.24)$$

должна приближенно подчиняться $F(m_1, m_2)$ -распределению с числом степеней свободы числителя $m_1 = s - k$ и знаменателя $m_2 = n - s$. Все величины в формуле (10.24) соответствуют ранее введенным обозначениям. В частности, X_i^0 — середина i -го гиперпараллелепипеда группирования (или i -е значение аргумента, в котором было проведено ν_i наблюдений); $f_a(X_i^0; \hat{\Theta})$ — значение гипотетической функции регрессии, вычисленное в точке $X = X_i^0$; $\bar{y}_i$ — условное среднее из ординат, попавших в i -й гиперпараллелепипед группирования (или из ординат, измеренных при i -м фиксированном значении аргумента X_i^0); y_{ij} — j -е по счету значение ординаты из числа попавших в i -й интервал группирования (или из числа измеренных при i -м фиксированном значении аргумента X_i^0). Легко понять, что числитель в правой части (10.24) характеризует меру рассеивания экспериментальных данных вокруг аппроксимирующей выборочной регрессионной поверхности, а знаменатель — меру рассеивания экспериментальных данных около своих условных выборочных средних $\bar{y}_i$ (т. е. меру, независимую от выбранного вида линии регрессии). Причем и числитель, и знаменатель являются практически независимыми (в некоторых частных случаях — *точно* независимыми) статистическими оценками одной и той же теоретической дисперсии $\sigma^2 = D(\eta | \xi = X)$.

Соответственно получаем следующее правило проверки гипотезы об общем виде функции регрессии. Задаемся, как обычно, достаточно малым уровнем значимости критерия α (например, $\alpha = 0,05$). С помощью таблицы находим $100(1 - \frac{\alpha}{2})\%$ -ную точку $v_{1-\frac{\alpha}{2}}^2$ и $100\frac{\alpha}{2}\%$ -ную точку $v_{\frac{\alpha}{2}}^2$ $F(k - m, n - k)$ -распределения. Если окажется, что величина v^2 , подсчитанная по формуле (10.24), удовлетворяет неравенствам

$$v_{1-\frac{\alpha}{2}}^2 < v^2 < v_{\frac{\alpha}{2}}^2,$$

то высказанная нами гипотеза об общем виде функции регрессии признается не противоречащей экспериментальным данным (10.1). Если же эти

неравенства оказались нарушенными, то гипотеза об общем виде функции регрессии отвергается с уровнем значимости α . При этом если v^2 «слишком мало» (т. е. $v^2 < v_{1-\frac{\alpha}{2}}^2$), то, очевидно, при выборе общего вида регрессии мы неправомерно реагировали на случайные отклонения точек $(X_i^0, \bar{y}_i)$ от истинной функции регрессии и тем самым необоснованно повысили число параметров k , от которых зависит уравнение регрессии. Напротив, если v^2 «слишком велико» (т. е. $v^2 > v_{\frac{\alpha}{2}}^2$), то «гибкость» аппроксимирующей функции регрессии $f_a(X; \Theta)$ следует признать недостаточной, поэтому целесообразно увеличить число неизвестных параметров регрессии (например, повысить порядок аппроксимирующего полинома).

Для случая, когда условная дисперсия зависимой переменной пропорциональна некоторой известной функции аргумента, т. е. $D\eta(X) = \sigma^2 h^2(X)$, формула (10.24) преобразуется:

$$v'^2 = \frac{\frac{1}{s-k} \sum_{i=1}^s w_i \nu_i |\bar{y}_i - f_a(X_i^0, \hat{\Theta})|}{\frac{1}{n-s} \sum_{i=1}^s w_i \sum_{j=1}^{\nu_i} (y_{ij} - \bar{y}_i)^2}, \quad (10.24')$$

где

$$w_i = 1/h^2(X_i^0).$$

Так, в примере 10.1: $n = 40$; $s = 4$; $\alpha = 0,05$; $k = 2$; дисперсионное отношение v'^2 , подсчитанное по формуле (10.24'), равно 1,04, в то время как 5%-ная точка $F(2,36)$ -распределения $v_{0,05}^2 = 3,26$. Это свидетельствует о том, что гипотеза о линейном виде регрессионной зависимости в данном случае не противоречит имеющимся в нашем распоряжении экспериментальным данным.

При проверке *линейности* регрессии (так же, впрочем, как и при проверке гипотезы о полиномиальном характере регрессии заданного порядка) в нормальных схемах зависимостей типа B и C_1 (см. п. 10.5) описанный общий критерий является *точным*. При этом в линейном случае статистика v^2 , определенная соотношением (10.24), может быть выражена в более удобной форме, не требующей предварительного вычисления выборочной аппроксимирующей функции регрессии. Так, например, в случае *парной* линейной регрессии (т. е. при $k = 2$) имеем:

$$v^2 = \frac{(n-s)(\hat{\rho}_{\eta,\xi}^2 - \hat{r}^2)}{(s-2)(1 - \hat{\rho}_{\eta,\xi}^2)}. \quad (10.25)$$

Здесь $\hat{\rho}_{\eta,\xi}$ и $\hat{r}$ — соответственно выборочные корреляционные отношения (η по ξ) и коэффициент корреляции (см. гл. 11). Логическая схема исполь-

зования статистики (10.25) аналогична ранее изложенным критериям: задаются достаточно малым ($0,05 \sim 0,15$) уровнем значимости α ; находят по таблице $100\alpha\%$ -ную точку v_α^2 распределения $F(s-2, n-s)$; сравнивают величину v^2 , определенную с помощью (10.25), с процентной точкой v_α^2 ; если оказывается, что $v^2 > v_\alpha^2$, то гипотезу о линейном виде регрессии считают статистически необоснованной.

Воспользуемся данным критерием для статистической проверки линейности регрессии в примере 10.2. Вычисления дают: $\bar{r}^2 = 0,429$, $\rho_{\eta, \xi}^2 = 0,459$, так что $v^2 = 0,513$. Принимая во внимание, что величина 5%-ной точки $F(4,37)$ -распределения равна $v_{0,05}^2 = 2,63$, делаем вывод о непротиворечивости гипотезы линейной регрессии и данных нашего эксперимента в данном примере ($0,513 < 2,63$).

2. *Общий приближенный критерий, основанный на негруппированных данных (при известной величине дисперсии остаточной случайной компоненты).*

Встречаются ситуации, когда в результате предварительных исследований или из других каких-либо соображений нам удастся заранее определить величину дисперсии σ^2 остаточной случайной компоненты ϵ в разложениях вида (10.14) и (10.16) (например, когда ϵ — ошибка измерения и нам известны характеристики точности используемого измерительного прибора). В этом случае можно отказаться от стеснительного требования группированности данных и для проверки гипотезы об общем виде функции регрессии воспользоваться фактом $\chi^2(n-k)$ -распределенности статистики

$$\gamma = \frac{1}{\sigma^2} \cdot \frac{1}{n-k} \sum_{i=1}^n (y_i - f_a(X_i; \hat{\Theta}))^2 \quad (10.26)$$

(который имеет место при условии справедливости нашей гипотезы).

Задавшись уровнем значимости критерия α и найдя с помощью табл. величины $100(1 - \frac{\alpha}{2})\%$ -ных и $100\frac{\alpha}{2}\%$ -ных точек χ^2 -распределения с $n-k$ степенями свободы соответственно $\chi_{1-\frac{\alpha}{2}}^2(n-k)$ и $\chi_{\frac{\alpha}{2}}^2(n-k)$, проверяем выполнение неравенства

$$\chi_{1-\frac{\alpha}{2}}^2(n-k) < \gamma < \chi_{\frac{\alpha}{2}}^2(n-k),$$

где γ подсчитано по формуле (10.26). Если эти неравенства оказались нарушенными, то от гипотезы H_0 об общем виде функции регрессии следует отказаться. При этом если γ «слишком мало» (т.е. $\gamma \leq \chi_{1-\frac{\alpha}{2}}^2(n-k)$), то, очевидно, при выборе общего вида мы неправильно реагировали на случайные отклонения экспериментальных точек (X_i, y_i) и тем самым необоснованно завысили число параметров k , от которых зависит уравнение

регрессии.

Напротив, если γ «слишком велико» (т. е. $\gamma \geq \chi^2_{\frac{\alpha}{2}}(n - k)$), то «гибкость» аппроксимирующей кривой регрессии $f_a(X; \hat{\Theta})$ следует признать недостаточной, поэтому целесообразно увеличить число неизвестных параметров регрессии (например, повысить порядок аппроксимирующего полинома).

Для случая, когда дисперсия зависимой переменной (или, что то же, дисперсия остаточной случайной компоненты) не остается постоянной при изменении X , а пропорциональна некоторой известной функции аргумента, т. е. $D\eta(X) = \sigma^2 h^2(X)$, формула подсчета статистики γ несколько изменится:

$$\gamma = \frac{1}{\sigma^2} \cdot \frac{1}{n - k} \sum_{i=1}^n w_i (y_i - f_a(X_i; \hat{\Theta}))^2,$$

где $w_i = 1/h^2(X_i)$. В остальном схема проверки гипотезы об общем виде функции регрессии остается той же самой, что и в случае $D\eta(X) = \sigma^2 = \text{const}$.

10.7.4. Некоторые общие рекомендации

При выборе общего вида искомой функции регрессии $f(X)$, помимо соображений и приемов, описанных выше, полезно учитывать следующие общие рекомендации.

1) *Не следует гнаться за чрезмерной сложностью функции, описывающей поведение искомой функции регрессии*, руководствуясь исключительно соображениями оптимизации критерия качества аппроксимизации $\Delta_n(f)$ (см. (10.21)). Дело в том, что если и оценки $\hat{\Theta}$ неизвестных параметров модели Θ , и значение критерия $\Delta_n(f)$ *вычисляются на основании одной и той же выборки*, то за счет увеличения размерности k оцениваемого векторного параметра $\Theta = (\theta_1, \theta_2, \dots, \theta_k)^T$ можно добиться, на первый взгляд, идеального результата. Возможность эта основана на известном в математическом анализе результате, в соответствии с которым для любой заданной системы точек $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ (с неповторяющимися абсциссами) можно подобрать алгебраический полином степени $n - 1$, проходящий в точности через все точки этой системы. А это значит, что все «невязки» $y_i - f(x_i)$, на основании которых строится критерий $\Delta_n(f)$ (а следовательно, и сам критерий, см. п. 10.6, этап 6), равны нулю, т. е. «лучше» модели для описания поведения функции регрессии подобрать невозможно. На самом деле при таком подходе мы как бы заставляем функцию $f(x)$ реагировать на случайные флуктуации, объясняемые наличием в представлениях типа (10.14) остаточной случайной компоненты

$\varepsilon(X)$. Поэтому, если мы попробуем применить полученный таким образом результат к другой выборке из той же самой генеральной совокупности, то увидим явное рассогласование модельных ($f(X_i)$) и наблюдаемых (y_i) значений результирующего показателя. Поэтому при подборе общего вида функции регрессии, как правило, *идут от простого к сложному*, т. е. начиная с анализа возможности использовать простейшую линейную модель вида (10.23).

2) *Следует добиваться компромисса между сложностью регрессионной модели и точностью ее оценивания.* Из общих результатов математической статистики, относящихся к анализу точности оценивания исследуемой модели при ограниченных объемах выборки, следует, что с увеличением сложности модели (выраженной, например, размерностью k векторного параметра Θ , участвующего в ее уравнении) точность оценивания падает. Например, ширина доверительного интервала $\Delta[y(X)]_P$ для неизвестного значения $y(X)$, при прочих равных характеристиках анализируемой схемы, увеличивается с ростом размерности параметра Θ , участвующего в вычислении функции регрессии $f(X; \Theta)$ (см. гл. 2 тома 2). Именно поэтому в ситуациях, когда исследователь располагает ограниченной исходной выборочной информацией вида (10.1), он вынужден искать компромисс между степенью общности привлекаемого класса допустимых решений F и точностью оценивания, которой возможно при этом добиться (пример подобного рода действий по поиску компромисса будет приведен в гл. 2 тома 2 в связи с задачей определения оптимального числа объясняющих переменных в линейной модели множественной регрессии).

3) *При обнаружении нелинейности в парных статистических связях анализируемых переменных $x^{(j)}$ и y ($j = 1, 2, \dots, p$) следует попытаться применить к этим переменным линеаризующие преобразования.* Простейший пример такого приема мы имеем, когда вместо анализа степенной зависимости вида

$$y = \theta_0(x)^{\theta_1} \quad (10.27)$$

исследователь рассматривает линейную зависимость между логарифмами исходных переменных, а именно:

$$\bar{y} = \bar{\theta}_0 + \theta_1 \bar{x}, \quad (10.28)$$

где $\bar{y} = \ln y$, $\bar{x} = \ln x$ и $\bar{\theta}_0 = \ln \theta_0$. В зависимости от типа нелинейной связи, существующей между исходными переменными, подбираются и другие линеаризующие преобразования (см., например, [Айвазян С. А., Енюков И. С., Мешалкин Л. Д., 1985, п. 6.2.3]). Один из наиболее общих подходов к линеаризации анализируемых зависимостей реализуется с по-

мощью так называемого *преобразования Бокса-Кокса* (см. гл. 2 в томе 2 учебника).

4) *Анализ регрессионных остатков.* Ряд статистических критериев проверки адекватности используемой аппроксимирующей модели регрессии основан на анализе регрессионных остатков (невязок) $\hat{\varepsilon}(X_i) = y_i - \hat{f}_a(X_i), i = 1, 2, \dots, n$. В основе их конструирования — положение, в соответствии с которым правильный выбор модели $f_a(X)$ предопределяет асимптотическую (по $n \rightarrow \infty$) независимость остатков $\hat{\varepsilon}(X_i)$. Поэтому статистическая проверка правильности выбора общего вида функции регрессии сводится к проверке статистической независимости остатков, для чего могут быть использованы, например, критерии, описанные в гл. 3 тома 2 (п. 3.2). В выражении для невязок $\hat{\varepsilon}(X_i)$ под $\hat{f}_a(X_i)$ понимается значение оцененной аппроксимирующей функции регрессии в точке X_i , т. е. $\hat{f}_a(X_i) = f_a(X_i; \hat{\Theta})$.

5) *Поиск модели, наиболее устойчивой к варьированию состава выборочных данных, на основании которых она оценивается.* Идея этого подхода к выбору общего вида исследуемой регрессионной зависимости основана на следующем простом соображении. Если общий параметрический вид зависимости $f(x^{(1)}, x^{(2)}, \dots, x^{(p)}; \Theta)$ «угадан» правильно, то результаты оценивания $\hat{\Theta}_1, \hat{\Theta}_2, \dots$ параметра Θ по различным подвыборкам выборки $\{x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)}; y_i\}_{i=1, \dots, n}$ будут мало отличаться друг от друга (а следовательно, не сильно будут различаться между собой и соответствующие значения $f(x^{(1)}, x^{(2)}, \dots, x^{(p)}; \hat{\Theta}_1), f(x^{(1)}, x^{(2)}, \dots, x^{(p)}; \hat{\Theta}_2), \dots$). И наоборот, при неудачном выборе общего вида искомой зависимости результаты ее восстановления по различным выборкам, как правило, будут сильно отличаться один от другого. Описание реализации этого подхода читатель может найти в [Айвазян С. А., Енюков И. С., Мешалкин Л. Д., 1985, п. 6.2.3].

ВЫВОДЫ

1. Аппарат статистического исследования зависимостей — составная часть многомерного статистического анализа — нацелен на решение основной проблемы естествознания: как на основании частных результатов статистического наблюдения за анализируемыми событиями или показателями выявить и описать существующие между ними стохастические взаимосвязи.

2. Анализируемые переменные величины по своей роли в исследовании подразделяются на результирующие (прогнозируемые) Y и объясня-

ющие (экзогенные, предикторные) X . Среди компонент векторов Y и X могут быть и количественные, и порядковые (ординальные), и классификационные (номинальные).

3. Центральным математическим объектом в процессе статистического исследования зависимостей является функция $f(X)$, называемая функцией регрессии Y по X и описывающая изменение условного среднего значения $Y_{\text{ср}}(X)$ результирующего показателя Y (вычисленного при фиксированных на уровне X значениях объясняющих переменных) в зависимости от изменения значений объясняющих переменных X .

4. Конечные прикладные цели статистического исследования зависимостей могут быть в основном трех типов: 1) установление самого факта наличия (или отсутствия) статистически значимой связи между Y и X , исследование структуры этих связей; 2) прогноз (восстановление) неизвестных значений индивидуальных или средних значений результирующего показателя по заданным значениям соответствующих объясняющих (экзогенных) переменных; 3) выявление причинных связей между объясняющими переменными X и результирующими показателями Y , частичное управление значениями Y путем регулирования величин объясняющих переменных X .

5. Разделы многомерного статистического анализа, составляющие математический аппарат статистического исследования зависимостей, формировались и развивались с учетом специфики анализируемых моделей, обусловленной в первую очередь природой исследуемых переменных. Так, изучение зависимостей между количественными переменными обслуживается регрессионным и корреляционным анализами и анализом временных рядов (гл. 11, а также 2, 3 и 4 тома 2); изучение зависимостей количественного результирующего показателя от неколичественных или разнотипных объясняющих переменных — дисперсионным и ковариационным анализами, моделями типологической регрессии; наконец, для исследования системы зависимостей, в которых одни и те же переменные в разных уравнениях этой системы могут одновременно выполнять и роль результирующих, и роль объясняющих, служит теория одновременных эконометрических уравнений (гл. 1 и 4 тома 2). Аппарат исследования зависимостей неколичественных или разнотипных результирующих показателей от количественных или разнотипных объясняющих переменных в книге не рассматривается.

6. К основным типовым задачам практики, в которых использование аппарата статистического исследования зависимостей оказывается наиболее уместным и эффективным, следует отнести задачи: 1) нормирования; 2) прогноза, планирования и диагностики; 3) оценки труднодоступных

(для непосредственного наблюдения и измерения) характеристик исследуемой системы; 4) оценки эффективности функционирования (или качества) анализируемой системы; 5) регулирования параметров функционирования анализируемой системы. Все эти задачи являются основными составными частями центральной проблемы кибернетики — проблемы «управления, связи и разработки информации» (см.: Математическая энциклопедия. Т. 2 — М.: Советская энциклопедия, 1979, с. 850).

7. По своей природе исследуемые зависимости могут быть разделены на: 1) детерминированные (тип А), когда исследуется функциональная зависимость между неслучайными переменными; 2) регрессионные (тип В), когда исследуется зависимость случайного результирующего показателя от неслучайных объясняющих переменных — параметров системы; 3) корреляционные (тип С), когда исследуется зависимость между случайными переменными, причем объясняющие переменные могут быть измерены без искажений; 4) конфлюэнтные (тип D), когда исследуется функциональная зависимость между случайными или неслучайными переменными в ситуации, когда те и другие могут быть измерены только с некоторой случайной ошибкой.

8. Весь процесс статистического исследования зависимостей может быть разбит на семь последовательно реализуемых основных этапов, хронологический характер связей которых дополняется связями итерационного взаимодействия (см. рис. 10.6): этап 1 (постановочный); этап 2 (информационный); этап 3 (корреляционный анализ); этап 4 (определение класса допустимых решений); этап 5 (анализ мультиколлинеарности предсказывающих переменных и отбор наиболее информативных из них); этап 6 (вычисление оценок неизвестных параметров, входящих в исследуемое уравнение статистической связи); этап 7 (анализ точности полученных уравнений связи).

9. *Этап параметризации регрессионной модели*, т. е. выбора параметрического семейства функций (класса допустимых решений), в рамках которого производится дальнейший поиск неизвестной функции регрессии, *является одновременно наиболее важным и наименее теоретически обоснованным* этапом регрессионного анализа.

10. Прежде всего исследователь должен сосредоточить свои усилия на *анализе содержательной сущности* искомой статистической зависимости, чтобы максимально использовать имеющиеся априорные сведения о «физическом» механизме изучаемой связи при выборе общего вида функции регрессии.

11. Важную роль в правильном выборе параметрического класса допустимых решений играет *предварительный анализ геометрической*

структуры совокупности исходных данных и в первую очередь анализ геометрии парных корреляционных полей, включающий в себя, в частности, учет и формализацию «гладких» свойств искомой функции регрессии, использование вспомогательных линеаризующих преобразований.

12. Сформулированные с помощью содержательного и геометрического анализа рабочие гипотезы об общем виде искомой функции регрессии могут быть проверены с привлечением соответствующих *математико-статистических критериев*. Среди фундаментальных идей, на которых базируются эти статистические критерии, следует выделить: а) идею компромисса между сложностью регрессионной модели («емкостью» класса допустимых решений) и точностью ее оценивания; б) идею поиска модели, наиболее устойчивой к варьированию состава выборочных данных, на основании которых она оценивается; в) идею проверки гипотез об общем виде функции регрессии на базе сравнения выборочных критериев адекватности и исследования статистических свойств получаемых при этом оценок размерности модели.

ГЛАВА 11. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ МНОГОМЕРНОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

11.1. Назначение и место корреляционного анализа в статистическом исследовании

При исследовании реальных социально-экономических явлений и систем статистику, экономисту приходится сталкиваться, как правило, с необходимостью статистического анализа *многомерной* генеральной совокупности, т.е. с ситуациями, когда на каждом из статистически обследуемых объектов этой совокупности регистрируются значения *целого набора* признаков $x^{(1)}, x^{(2)}, \dots, x^{(p)}$. В дальнейшем мы будем обозначать, как и прежде, этот набор признаков с помощью $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^T$, а результат регистрации значений этих признаков на i -м статистически обследованном объекте — i -е *многомерное наблюдение* — с помощью $X_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)})^T$. Таким образом, «стартовая позиция» при статистическом анализе многомерной генеральной совокупности аналогична одномерному случаю: *исследователь по имеющейся у него случайной выборке*

$$X_1, X_2, \dots, X_n \quad (11.1)$$

значений анализируемой многомерной случайной величины $\xi = (\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(p)})^T$ должен сделать те или иные статистические выводы о ее «поведении».

Мы уже знаем (см. п. 2.6), что *исчерпывающие* сведения о поведении анализируемой случайной величины ξ содержатся в ее законе распределения вероятностей (з.р.в.) $f_\xi(X)$, где под $f_\xi(X)$ понимается значение функции плотности вероятности в точке X , если случайная величина ξ непрерывна, и вероятность $P\{\xi = X\}$ того, что случайная величина

примет значение, равное X , если ξ дискретна. Однако з.р.в. анализируемой случайной величины исследователю, как правило, неизвестен. И если при описании поведения одномерных случайных величин исследователь еще имеет *практически реализуемые* возможности подбора и использования подходящих модельных законов распределения (см. гл. 3) с последующей статистической оценкой участвующих в их записи параметров (см. гл. 7), то при исследовании признаков размерности $p \geq 2$ ему чаще всего приходится ограничиваться информацией, которую доставляют оценки моментов первых двух порядков, а именно оценками *вектора средних значений* $\widehat{M}_1 = (\hat{m}_1^{(1)}, \hat{m}_1^{(2)}, \dots, \hat{m}_1^{(p)})^T$ и *ковариационной матрицы* $\widehat{\Sigma} = (\hat{\sigma}_{ij})$, $-i, j = 1, 2, \dots, p$ (см. п. 6.2). Другими словами, можно сказать, что за редчайшим исключением¹, все выводы *многомерного статистического анализа строятся на базе оценок* $\widehat{M}_1$ и $\widehat{\Sigma}$.

Оценка $\widehat{M}_1$ вектора средних значений дает представление о *центре группирования* наблюдений анализируемого многомерного признака. Отличие от одномерного случая, когда мы располагаем оценкой $\hat{m}_1 = \bar{x}$ среднего значения анализируемой случайной величины, заключается лишь в том, что $\widehat{M}_1$ определяет точку в p -мерном пространстве, в то время как выборочное среднее $\hat{m}_1 = \bar{x}$ определяет точку на *числовой прямой*. По существу вся специфика многомерного случая сосредоточена в ковариационной матрице Σ , а при статистическом анализе — в ее оценке $\widehat{\Sigma}$. Именно знание ковариационной матрицы позволяет исследователю строить и анализировать характеристики случайного рассеивания и статистической взаимосвязи (*коррелированности*) компонент анализируемого многомерного признака. Данная глава как раз и посвящена так называемому **корреляционному анализу** многомерной генеральной совокупности, назначение которого — получить (на основе имеющейся выборки (11.1)) ответы на следующие основные вопросы:

- как выбрать (с учетом специфики и природы анализируемых переменных) подходящий измеритель статистической связи (коэффициент корреляции, корреляционное отношение, какую-либо информационную характеристику связи, ранговый коэффициент корреляции и т. п.)?
- как оценить (с помощью точечной и интервальной оценок) его числовое значение по имеющимся выборочным данным?
- как проверить гипотезу о том, что полученное числовое значение анализируемого измерителя связи действительно свидетельствует

¹ Исключения относятся в основном к генеральным совокупностям, описываемым либо многомерным нормальным, либо полиномиальным з.р.в. (см. соответственно (3.9') и (3.7)).

о наличии статистической связи (или, как говорят, проверить исследуемую корреляционную характеристику на статистически значимое ее отличие от нуля)?

• *как определить структуру связей между компонентами исследуемого многомерного признака, сопоставив каждой паре компонент двочный ответ («связь есть» или «связи нет»)?*

Какое место занимает корреляционный анализ в общем прикладном статистическом исследовании? В гл. 9 обозначены три центральные проблемы прикладной статистики. Начнем с первой и наиболее значимой из них — с проблемы *статистического исследования зависимостей*. В соответствии с предложенным в предыдущей (10-й) главе разбиением процесса решения этой проблемы на этапы (см. п. 10.6) корреляционный анализ составляет содержание, по существу, *первого* (после постановки задачи и сбора необходимых статистических данных) этапа статистического исследования зависимостей. Действительно, при исследовании зависимостей между анализируемыми переменными мы должны дать в первую очередь ответ на вопрос: *а существует ли такая зависимость или анализируемые признаки статистически независимы?* И только после утвердительного ответа на этот вопрос заняться выявлением вида и математической формы этой зависимости. Корреляционный анализ как раз и предоставляет средства, позволяющие ответить на первый вопрос. Выявление же вида и математической формы искомой зависимости производится с помощью разнообразных методов и моделей регрессионного анализа и анализа временных рядов, которым посвящен том 2 учебника. В то же время рассматриваемые в рамках корреляционного анализа характеристики статистической связи (ковариации, различные коэффициенты корреляции и т. п.) используются в качестве «входной» (базовой) информации при решении задач других двух центральных проблем прикладной статистики — *классификации объектов и признаков* и *снижения размерности анализируемого признакового пространства* (методы и модели, предназначенные для решения этих двух проблем, описываются в гл. 12 и 13). Поэтому именно с главы, посвященной корреляционному анализу, по существу, начинается изложение методов и моделей прикладной статистики и эконометрики.

11.2. Корреляционный анализ количественных признаков

Здесь речь идет об измерителях степени тесноты статистической связи между *количественными* компонентами вектора $X = (x^{(1)}, x^{(2)}, \dots)$,

$\dots, x^{(p)}\rangle^T$. Напомним (см. п. 2.3), что значение количественной случайной величины есть результат измерения степени проявления анализируемого свойства статистически обследованного объекта в числовой шкале определенного физического смысла (в штуках, метрах, денежных единицах, килограммах, единицах времени и т. д.).

В соответствии с изложенной в п. 10.1 общей логической схемой статистического исследования зависимостей анализируемые переменные по своей роли делятся на *результатирующие* и *объясняющие*. Поэтому, оставив за компонентами вектора X роль объясняющих переменных, введем в рассмотрение *результатирующую переменную* y . Тогда исходные статистические данные (11.1) пополнятся еще одним вектором-столбцом $Y = (y_1, y_2, \dots, y_n)^T$, так что в дальнейшем предметом нашего анализа будут исходные статистические данные вида

$$X^{(1)}, X^{(2)}, \dots, X^{(p)}, Y, \quad (11.1')$$

где $X^{(j)} = (x_1^{(j)}, x_2^{(j)}, \dots, x_n^{(j)})^T$ — вектор-столбец наблюдаемых значений j -й объясняющей переменной ($j = 1, 2, \dots, p$).

В п. 10.5 описаны основные типы статистических зависимостей (B, C и D), которые могут наблюдаться между исследуемыми переменными. Умение правильно классифицировать каждую конкретную многомерную систему наблюдений играет решающую роль при правильном выборе подходящих математико-статистических методов анализа изучаемой зависимости на *стадии регрессионного анализа и при неформальной содержательной интерпретации выявленной связи*.

Однако при решении задач корреляционного анализа удобнее использовать *единый (унифицированный) подход*, при котором исследуемая *объясняющая переменная* (*случайная* при типах зависимостей C и D и *неслучайная* при зависимостях типа B) *интерпретируется как параметр, от которого зависит закон распределения результирующего показателя*. Его мы и будем придерживаться в данной главе.

11.2.1. Коэффициент детерминации как универсальная характеристика степени тесноты статистической связи

Основная идея, лежащая в основе определения коэффициента детерминации, состоит в следующем. Пусть нас интересует *степень тесноты статистической связи* (с.т.с.с.), существующей между результирующим показателем y и объясняющей переменной (вообще говоря, векторной) X . Очевидно, степень тесноты этой связи может считаться *максимальной*,

если по заданному значению объясняющей переменной X можно *однозначно* (без всякой случайной ошибки) восстановить соответствующее значение результирующего показателя $y(X)$. И обратно: *если значение величин X не несет никакой информации о значении результирующего показателя y* , то связь отсутствует вовсе, и соответствующий измеритель степени ее тесноты должен принимать *минимальное* (на выбранной шкале) возможное значение.

Для того чтобы математически формализовать эту естественную посылку в виде подходящего измерителя с.т.с.с., посмотрим, как она интерпретируется в терминах регрессионной модели типа (10.14).

Итак, мы рассматриваем статистическую зависимость вида

$$y(X) = f(X) + \varepsilon(X), \quad (11.2)$$

где $f(X) = E(y | X)$ — условное среднее значение результирующего показателя (при условии, указанном за прямой чертой, здесь и далее означает, что объясняющая переменная приняла заданное *фиксированное* значение X), т.е. *функция регрессии y по X* , — а $\varepsilon(X)$ остаточная случайная компонента, з.р.в. которой, вообще говоря, может зависеть от X , но при этом предполагается:

$$\begin{aligned} E(\varepsilon | X) &\equiv 0^1, \\ D(\varepsilon | X) &= \sigma_\varepsilon^2(X), \end{aligned} \quad (11.2')$$

и $\varepsilon(X)$ не коррелирована с $f(X)$ в схемах зависимости типов C и D (в схемах типа B величина $f(X)$ не является случайной).

Отметим два важных соотношения, связывающих характеристики варьирования (условные и безусловные) составных частей модели (11.2)–(11.2'):

$$E(y - y_0)^2 = E(f - f_0)^2 + E\varepsilon^2, \quad (11.3)$$

$$D(y | X) = D(\varepsilon | X) = \sigma_\varepsilon^2(X) \quad (11.4)$$

В соотношении (11.3), связывающем между собой безусловные характеристики варьирования, $y_0 = Ey$, $f_0 = Ef$, а символ E означает *усреднение следующих за ним величин по всем возможным значениям как y , так и X* . Нуждается в пояснении случай зависимости по схеме B (см. п.10.5), в которой объясняющая переменная X не является случайной. В этом случае усреднение по X производится по всем наблюдаемым значениям $X_1, X_2, \dots, X_n$, так что, например:

$$f_0 = Ey = \frac{1}{n} \sum_{i=1}^n f(X_i) = \frac{1}{n} \sum_{j=1}^n \nu_j f(X_j^0);$$

$$\sigma_f^2 = E(f - f_0)^2 = \frac{1}{n} \sum_{i=1}^n (f(X_i) - f_0)^2 = \frac{1}{n} \sum_{j=1}^n \nu_j (f(X_j^0) - f_0)^2;$$

¹ Знак тождества « $\equiv$ » означает, что среднее значение ε равно нулю при всех возможных значениях объясняющей переменной X .

$$E\varepsilon^2 = \frac{1}{n} \sum_{i=1}^n E\varepsilon^2(X_i) = \frac{1}{n} \sum_{j=1}^s \nu_j E\varepsilon^2(X_j^0)$$

(вариант формул для группированных данных дается в обозначениях п. 6.1). Что касается условных характеристик, участвующих в (11.4), то они были определены выше (здесь усреднение производится по всем возможным значениям y при заданном фиксированном значении X).

Вернемся непосредственно к обсуждению возможного способа измерения с.т.с.с. Из (11.2) видно, что однозначное восстановление y по X возможно только в ситуации «вырождения» (отсутствия) случайной остаточной компоненты $\varepsilon(X)$ при всех возможных значениях X , что равносильно (при нулевом среднем значении $\varepsilon(X)$) отсутствию какого-либо разброса значений $\varepsilon(X)$ относительно нуля. А это означает на языке вероятностных характеристик тождественное равенство нулю условных дисперсий $\varepsilon(X)$ и $y(X)$, т. е.

$$D(\varepsilon | X) = D(y | X) = \sigma_\varepsilon^2(X) \equiv 0. \quad (11.5)$$

Другой крайний (с точки зрения с.т.с.с.) случай — это ситуация, в которой знание величины объясняющей переменной X не имеет никакой информации о значении результирующего показателя y . Эта ситуация возможна, в рамках модели (11.2), лишь тогда (и только тогда), когда

$$f(X) \equiv c = \text{const}. \quad (11.6)$$

В этом случае поведение результирующего показателя y (определенного соотношением (11.2) с учетом (11.6)) как бы повторяет, с точностью до сдвига на постоянную величину c , поведение остаточной случайной компоненты ε и, в частности, в силу (11.3) и с учетом тождества $E(f - f_0)^2 \equiv 0$ имеет в точности ту же самую меру случайного рассеивания, т. е.

$$E(y - y_0)^2 = E\varepsilon^2$$

или

$$Dy = D\varepsilon. \quad (11.7)$$

Теперь мы подготовлены к тому, чтобы ввести и проинтерпретировать достаточно универсальный измеритель с.т.с.с. между y и X — коэффициент детерминации y по X . Обозначим его с помощью $K_d(y; X)$ и определим соотношением

$$K_d(y; X) = 1 - \frac{D\varepsilon}{Dy}, \quad (11.8)$$

где $D\varepsilon = E\varepsilon^2$ — безусловная дисперсия остаточной случайной компоненты, которая может быть подсчитана как усредненная по всем возможным значениям X условная дисперсия $D(\varepsilon | X)$, т. е.

$$D\varepsilon = E_x(D(\varepsilon | X)) \quad (11.9)$$

(нижний индекс x у символа математического ожидания E означает, что усреднение проводится по всем возможным значениям X), а Dy — безусловная дисперсия результирующего показателя y , определенная соотношением (11.3).

Легко видеть, что введенный соотношением (11.8) коэффициент детерминации определяет в качестве шкалы измерения показателя с.т.с.с. отрезок $[0; 1]$, причем минимальное (нулевое) значение $K_d(y; X)$ соответствует полному отсутствию какой бы то ни было связи между y и X (поскольку в этом случае, как мы видели, см. (11.7), $D\varepsilon = Dy$), а максимальное значение $K_d(y; X) = 1$ соответствует случаю чисто функциональной зависимости между y и X , когда значение y может быть в точности восстановлено по значению X с помощью формулы $y = f(X)$ (так как $D\varepsilon$, определенное формулой (11.9) с учетом тождества (11.5), равно нулю). При этом из определения коэффициента детерминации (11.8) с учетом соотношения (11.3) следует, что численное значение $K_d(y; x)$ отражает долю общей вариации результирующего признака y , объясненную изменением функции регрессии $f(x)$.

До сих пор мы оперировали теоретическими характеристиками анализируемых переменных. Однако в статистическом анализе их заменяют соответствующими выборочными характеристиками. Вычисление выборочного (эмпирического) значения коэффициента детерминации y по X должно производиться по формуле

$$\hat{K}_d(y; X) = 1 - \frac{s_\varepsilon^2}{s_y^2}, \quad (11.8')$$

где

$$s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / n, \quad \bar{y} = \left(\sum_{i=1}^n y_i \right) / n,$$

а выборочное значение дисперсии «невязок» ε вычисляется по одной из двух формул:

$$s_\varepsilon^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(X_i))^2, \quad (11.9')$$

если предварительный анализ показал, что условная дисперсия $D(\varepsilon | X)$ не зависит от X (т.е. $D(\varepsilon | X) = \sigma_\varepsilon^2 = \text{const}$);

$$s_\varepsilon^2 = \frac{1}{n} \sum_{j=1}^s \sum_{i=1}^{\nu_j} (y_{ji} - \bar{y}_j)^2, \quad (11.9'')$$

если условная дисперсия $D(\varepsilon | X)$ зависит от значения объясняющей переменной X , или во всех случаях, когда вычисления ведутся по *группированной* выборке.

В соотношении (11.9') $\hat{f}(X_i)$ есть статистически оцененное значение неизвестной функции регрессии $f(X)$ в точке $X = X_i$ (методы оценивания функции регрессии описаны в гл. 2 тома 2). В соотношении (11.9'') использованы *группированные* выборочные данные и обозначения п. 6.1. В частности, s — это общее число интервалов группирования, ν_j — число выборочных данных, попавших в j -й интервал группирования, y_{ji} — значение результирующего признака в i -м по счету наблюдении, принадлежащем j -му интервалу группирования (нумерация наблюдений своя для каждого интервала группирования), а $\bar{y}_j$ — среднее значение результирующего признака, подсчитанное по наблюдениям, попавшим в j -й интервал группирования, т.е. $\bar{y}_j = (\sum_{i=1}^{\nu_j} y_{ji})/\nu_j$.

Пример 11.1. Подсчитаем коэффициент детерминации $\hat{K}_d(y, X)$ по данным примера 10.1 (см. табл. 10.1 и рис. 10.2). Вычисления по формулам (11.8'), (11.9') и (11.9'') дают:

$$\begin{aligned} \bar{y} &= 50,8; \quad s_y^2 = \frac{1}{40} \sum_{i=1}^{40} (y_i - \bar{y})^2 = 979,17; \\ s_\varepsilon^2 &= \frac{1}{40} \sum_{j=1}^4 9s^2(x_j^0) \\ &= \frac{9}{40} (40,96 + 256,00 + 510,76 + 835,21) = 369,96; \\ \hat{K}_d(y; x) &= 1 - \frac{369,96}{979,17} = 1 - 0,377 = 0,623. \end{aligned}$$

Такое значение коэффициента детерминации свидетельствует о наличии статистической связи между среднедушевыми денежными сбережениями семей y (η — в обозначениях примера 10.1) и их среднедушевым доходом x (ξ — в обозначениях примера 10.1), причем степень тесноты этой связи несколько выше среднего уровня. Более подробное обсуждение шкалы значений $\hat{K}_d$, так же как и вопросы *проверки этих значений на статисти-*

стически значимое отличие от нуля, отложим до описания некоторых частных версий этого важного индикатора.

11.2.2. Исследование линейной зависимости y от единственной объясняющей переменной x : **парный коэффициент корреляции**

Рассмотрение этой схемы зависимости между y и x разобьем на два случая: 1) переменные (x, y) «ведут себя» как двумерная нормальная случайная величина; 2) переменные x и y связаны *линейной* регрессионной зависимостью типа (11.2)-(11.2'), т. е. $y(x) = \theta_0 + \theta_1 x + \varepsilon(x)$, где θ_0 и θ_1 — некоторые неизвестные параметры, $\varepsilon(x)$ и x взаимно некоррелированы, а з.р.в. случайных величин y, ε и переменной x не обязан быть нормальным.

1. Случай двумерной нормальности случайных величин (x, y) . Эта схема зависимости относится к типу *C* по классификации, описанной в п. 10.5. Займемся анализом двумерного нормального распределения (см. п. 3.1.5, формулу (3.9)) с целью конкретизации вида функции регрессии $f(x)$ и других вероятностных характеристик модели (11.2). Понимая под $\xi^{(1)}$ и $\xi^{(2)}$ из формулы (3.9) соответственно объясняющую переменную x и результирующий показатель y и воспользовавшись результатами п. 3.1.5, имеем:

а)

$$f(x) = E(y | x) = E(\xi^{(2)} | \xi^{(1)} = x) = \theta_0 + \theta_1 x, \quad (11.10)$$

где

$$\theta_1 = r \frac{\sigma_{22}^{1/2}}{\sigma_{11}^{1/2}}, \quad \theta_0 = m_1^{(2)} - \theta_1 m_1^{(1)}, \quad (11.11)$$

а

$$\begin{aligned} m_1^{(1)} &= Ex, & \sigma_{11} &= \sigma_x^2 = E(x - m_1^{(1)})^2; \\ m_1^{(2)} &= Ey, & \sigma_{22} &= \sigma_y^2 = E(y - m_1^{(2)})^2; \end{aligned} \quad (11.12)$$

$$r(x, y) = \frac{E[(x - m_1^{(1)})(y - m_1^{(2)})]}{\sigma_x \sigma_y} = \quad (11.13)$$

так называемый **парный коэффициент корреляции** между переменными x и y ;

б)

$$D(y | x) = D(\xi^{(2)} | \xi^{(1)} = x) = \sigma_y^2(1 - r^2). \quad (11.14)$$

Из (11.10) следует, что если анализируемые переменные подчинены двумерному нормальному закону, то функция регрессии одной из них по

другой имеет линейный вид. Кроме того, из определения коэффициента детерминации (11.8) с учетом соотношений (11.9), (11.4) и (11.14) следует, что парный коэффициент корреляции r и коэффициент детерминации $K_d(y, x)$ связаны простым соотношением

$$K_d(y; x) = r^2(x, y). \quad (11.15)$$

При этом в отличие от коэффициента детерминации (который по определению может быть только неотрицательным числом) коэффициент корреляции может быть как положительным, так и отрицательным. А поскольку он имеет одинаковый знак с угловым коэффициентом наклона θ_1 прямой $f(x) = \theta_0 + \theta_1 x$ (см. соотношение (11.11)), то *положительное значение r свидетельствует о положительном (монотонно возрастающем) характере парной связи y и x , а отрицательное значение r — об отрицательном (монотонно убывающем) характере этой связи.*

Сформулируем и докажем основные свойства парного коэффициента корреляции (справедливые в рамках рассматриваемой двумерной нормальной генеральной совокупности).

С в о й с т в о 1. Шкала возможных значений парного коэффициента корреляции r ограничена отрезком от -1 до $+1$, т. е. $-1 \leq r \leq 1$, или $|r| \leq 1$.

Для доказательства этого свойства воспользуемся очевидными неравенствами

$$E \left(\frac{x - m_1^{(1)}}{\sigma_x} \pm \frac{y - m_1^{(2)}}{\sigma_y} \right)^2 \geq 0. \quad (11.16)$$

Возводя в квадрат *сумму* в круглых скобках, имеем

$$E \left(\frac{x - m_1^{(1)}}{\sigma_x} \right)^2 + 2E \left[\frac{(x - m_1^{(1)})(y - m_1^{(2)})}{\sigma_x \sigma_y} \right] + E \left(\frac{y - m_1^{(2)}}{\sigma_y} \right)^2 \geq 0. \quad (11.17)$$

Поскольку

$$E \left(\frac{x - m_1^{(1)}}{\sigma_x} \right)^2 = E(x - m_1^{(1)})^2 / \sigma_x^2 = 1$$

и

$$E \left(\frac{y - m_1^{(2)}}{\sigma_y} \right)^2 = E(y - m_1^{(2)})^2 / \sigma_y^2 = 1,$$

а

$$E \left[\frac{(x - m_1^{(1)})(y - m_1^{(2)})}{\sigma_x \sigma_y} \right] = r \quad (\text{по определению}),$$

то из (11.17) непосредственно следует, что $r \geq -1$. Аналогично рассуждая при анализе выражения, получающегося от возведения в квадрат *разности* в круглых скобках левой части выражения (11.16), получаем, что $r \leq 1$. Объединение полученных двух неравенств дает нам доказательство свойства 1.

Свойство 2. Если случайные величины x и y статистически независимы, то $r(x, y) = 0$.

Доказательство этого свойства следует из определения (11.13) с учетом того, что для статистически независимых случайных величин их ковариация равна нулю, т. е.

$$\begin{aligned} E[(x - m_1^{(1)})(y - m_1^{(2)})] &= [E(x - m_1^{(1)})][E(y - m_1^{(2)})] \\ &= (m_1^{(1)} - m_1^{(1)})(m_1^{(2)} - m_1^{(2)}) = 0. \end{aligned}$$

Свойство 3. Из факта $|r| = 1$ следует наличие чисто функциональной линейной связи между x и y и наоборот: если x и y связаны чисто функциональной линейной связью, то $|r| = 1$.

Для доказательства этого свойства следует обратить внимание на то, что неравенства $r \leq 1$ и $r \geq -1$ обращаются в точные равенства тогда и только тогда, когда обращаются в точные равенства неравенства (11.16) (по построению). А неравенства (11.16) обращаются в точные равенства тогда и только тогда, когда выражение в круглых скобках левой части этого соотношения тождественно равно нулю. Но из равенства

$$\frac{x - m_1^{(1)}}{\sigma_x} \pm \frac{y - m_1^{(2)}}{\sigma_y} = 0$$

следует, что x и y связаны линейной зависимостью.

Свойство 4. Коэффициент корреляции является *симметричной* характеристикой с.т.с.с. между x и y , т. е. $r(x, y) = r(y, x)$.

Доказательство этого свойства следует непосредственно из определения (11.13).

Свойство 5. Из равенства нулю коэффициента корреляции (т. е. из того, что $r(x, y) = 0$) следует статистическая независимость переменных x и y .

Чтобы доказать это свойство, надо показать, что из $r(x, y) = 0$ следует представление двумерной совместной плотности $f_{(x,y)}(u, v)$ в виде

произведения частных плотностей $f_x(u)f_y(v)$. Для этого воспользуемся общей формулой двумерного нормального распределения (3.9), подставив в нее значение $r = 0$. Получаем (с учетом введенных выше преобразований):

$$\begin{aligned} f_{(x,y)}(u,v) = \varphi(u,v) &= \frac{1}{2\pi\sigma_{11}^{1/2}\sigma_{22}^{1/2}} e^{-\frac{1}{2}\left[\frac{(u-m_1^{(1)})^2}{\sigma_{11}} + \frac{(v-m_1^{(2)})^2}{\sigma_{22}}\right]} \\ &= \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(u-m_1^{(1)})^2}{2\sigma_x^2}} \times \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(v-m_1^{(2)})^2}{2\sigma_y^2}}. \end{aligned} \quad (11.18)$$

Но правая часть (11.18) представляет собой как раз произведение частной плотности распределения вероятностей случайной величины x на частную плотность распределения вероятностей случайной величины y , что и завершает доказательство.

Дав определение и познакомившись с основными свойствами парного коэффициента корреляции r , мы должны привести формулу, по которой вычисляется эмпирический (выборочный) аналог этой характеристики — $\hat{r}(x, y)$:

$$\hat{r}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (11.13')$$

Однако для проведения полноценного статистического анализа свойств исследуемой генеральной совокупности, основанных на этой характеристике, *нам необходимо знать ее статистические свойства*. Это позволит судить о точности приближения (11.13') к неизвестному истинному значению (11.13), строить статистические критерии для проверки различных гипотез о численных значениях анализируемого коэффициента корреляции.

В частности, *какую величину выборочного коэффициента корреляции следует считать достаточной для статистически обоснованного вывода о наличии корреляционной связи между исследуемыми переменными?* Ведь надежность статистических характеристик, в том числе и $\hat{r}$, ослабевает с уменьшением объема соответствующей выборки, а потому принципиально возможны случаи, когда отклонение от нуля полученной величины выборочного коэффициента корреляции $\hat{r}$ оказывается *статистически незначимым*, т. е. целиком обусловленным неизбежным случайным колебанием выборки, на основании которой он вычислен. Ответить на этот вопрос помогает знание закона вероятностного распределения $\hat{r}$.

В случае совместной нормальной распределенности исследуемых переменных и при достаточно большом объеме выборки n (а именно при $n > 200$) распределение $\hat{r}$ можно считать приближенно нормальным со средним, равным своему теоретическому значению r , и дисперсией $\sigma_r^2 = \frac{(1-r^2)^2}{n}$.

Однако следует учитывать, что при малых значениях n и r , близких к ± 1 , это приближение оказывается очень грубым. Кроме того, при малых n следует принимать во внимание, что величина $\hat{r}$ является *смещенной* оценкой своего теоретического значения r , в частности $E\hat{r} = r - [r(1 - r^2)]/2n$.

Относительно хорошая степень приближения нормального распределения при малых значениях $|r|$ позволяет получить простой критерий проверки гипотезы

$$H_0: r = 0, \quad (11.19)$$

т. е. гипотезы об отсутствии корреляционной связи между исследуемыми переменными x и y . При этом используется тот факт, что статистика

$$\gamma_n = \frac{\hat{r}\sqrt{n-2}}{\sqrt{1-\hat{r}^2}}$$

при условии справедливости гипотезы H_0 приблизительно распределена по з.р.в. Стьюдента с $n - 2$ степенями свободы.

Поэтому если окажется, что

$$\frac{|\hat{r}|\sqrt{n-2}}{\sqrt{1-\hat{r}^2}} > t_{\frac{\alpha}{2}}(n-2), \quad (11.20)$$

то гипотеза H_0 об отсутствии корреляционной связи между x и y отвергается с вероятностью ошибиться, равной α (здесь, как и ранее, $t_q(m)$ — 100 q %-ная точка t -распределения с m степенями свободы).

Доверительные интервалы для истинного значения коэффициента корреляции r можно построить, используя следующее преобразование, предложенное Р. Фишером:

$$z = \frac{1}{2} \ln \frac{1+\hat{r}}{1-\hat{r}}. \quad (11.21)$$

Он показал, что величина z , определенная соотношением (11.21), уже при небольших n с хорошим приближением следует нормальному закону со средним $Ez \approx \frac{1}{2} \ln \frac{1+r}{1-r} + \frac{r}{2(n-1)}$ и дисперсией $Dz = \frac{1}{n-3}$. Это позволяет

построить доверительный интервал $[z_1, z_2]$ для $E\hat{r}$ по формуле

$$\begin{aligned} z_{1,2} &= \frac{1}{2} \ln \frac{1+\hat{r}}{1-\hat{r}} \mp \frac{u_{\frac{\alpha}{2}}}{\sqrt{n-3}} - \frac{\hat{r}}{2(n-1)} \\ &= \operatorname{arcth} \hat{r} \mp \frac{u_{\frac{\alpha}{2}}}{\sqrt{n-3}} - \frac{\hat{r}}{2(n-1)}, \end{aligned}$$

откуда следует, что истинное значение коэффициента корреляции r с той же доверительной вероятностью $1 - \alpha$ заключено в пределах

$$\operatorname{th} z_1 < r < \operatorname{th} z_2. \quad (11.22)$$

Здесь $\operatorname{th} z$ — это тангенс гиперболический от аргумента z (определяется с помощью соотношения $\operatorname{th} z = (e^z - e^{-z}) / (e^z + e^{-z})$). Соответственно функция, определяющая величину z с помощью соотношения (11.21), — это функция, обратная к тангенсу гиперболическому; так что часто вместо $z = \frac{1}{2} \ln \frac{1+\hat{r}}{1-\hat{r}}$ пишут $z = \operatorname{arcth} \hat{r}$ (или $z = \operatorname{th}^{-1} \hat{r}$). Нахождение z по данному значению $\hat{r}$ и, наоборот, определение $\hat{r}$ по заданной величине z производятся с помощью специальной табл. (см. приложение 1), в которой в крайних столбцах (левом и правом) приведены значения $|r|$, а между ними — соответствующие значения $|z| = \operatorname{arcth} |\hat{r}|$ (знаки у аргумента и функции совпадают, так что если, например, $\hat{r}$ отрицателен, то и соответствующее значение $z = \operatorname{arcth} \hat{r}$ также отрицательно).

Пример 11.2. По данным $n = 39$ предприятий получен коэффициент корреляции $\hat{r} = -0,654$, характеризующий тесноту связи между себестоимостью продукции (y) и производительностью труда (x). Построим интервальную оценку для r , задавшись 95%-й доверительной вероятностью.

По табл. из приложения 1 для $\hat{r} = -0,654$ найдем $z = -0,7823$. Тогда

$$\begin{aligned} z_1 &= -0,7823 - \frac{1,96}{\sqrt{36}} = -1,1090; \\ z_2 &= -0,7823 + \frac{1,96}{\sqrt{36}} = -0,4556. \end{aligned}$$

Теперь по табл. по найденным z_1 и z_2 найдем соответствующие значения $r_1 = -0,804$ и $r_2 = -0,426$.

Таким образом, можно утверждать, что с доверительной вероятностью $P = 0,95$ истинное значение коэффициента корреляции r между себестоимостью продукции y и производительностью труда x будет лежать в интервале от $-0,804$ до $-0,426$, т. е. $-0,804 \leq r \leq -0,426$.

Влияние ошибок измерения анализируемых переменных на величину коэффициента корреляции r . Пусть мы хотим оценить степень тесноты корреляционной связи между компонентами двумерной нормальной случайной величины (x, y) , однако наблюдать мы их можем лишь с некоторыми случайными «ошибками измерения» соответственно ε_x и ε_y (см. схему зависимости D в п. 10.5). Поэтому экспериментальные данные (x_i, y_i) , $i = 1, 2, \dots, n$, — это практически выборочные значения искаженной двумерной случайной величины (x', y') , где $x' = x + \varepsilon_x$ и $y' = y + \varepsilon_y$. Если предположить, что ε_x и ε_y взаимно независимы, не зависят от x и y , нормальны, имеют нулевые математические ожидания и конечные дисперсии, соответственно σ_1^2 и σ_2^2 , то двумерная случайная величина (x', y') будет также подчиняться двумерному нормальному распределению. Однако, как легко подсчитать, параметры этого распределения и, в частности, коэффициент корреляции r' между x' и y' будут соответственно отличаться от параметров исходной двумерной схемы (x, y) . Действительно, в соответствии с основными правилами вычисления первых и вторых моментов (см. п. 2.6) получаем

$$\begin{aligned} a_{x'} &= Ex' = Ex = a_x; \\ a_{y'} &= Ey' = Ey = a_y; \\ \sigma_{x'}^2 &= \sigma_x^2 + \sigma_1^2; \\ \sigma_{y'}^2 &= \sigma_y^2 + \sigma_2^2; \\ r' &= \frac{r}{\sqrt{\left(1 + \frac{\sigma_1^2}{\sigma_x^2}\right) \left(1 + \frac{\sigma_2^2}{\sigma_y^2}\right)}}. \end{aligned} \quad (11.23)$$

Из (11.23), в частности, следует, что коэффициент корреляции признаков, на которые наложены ошибки измерения, всегда меньше по абсолютной величине, чем коэффициент корреляции исходных признаков. Другими словами, ошибки измерения всегда ослабляют исследуемую корреляционную связь между исходными переменными, и это искажение тем меньше, чем меньше отношения дисперсий ошибок к дисперсиям самих исходных переменных. Формула (11.23) позволяет скорректировать искаженное значение коэффициента корреляции: для этого нужно либо знать «разрешающие» характеристики погрешностей измерений (и, следовательно, величины дисперсий ошибок σ_1^2 и σ_2^2), либо провести дополнительное исследование по их выявлению.

2. Общий случай парной линейной зависимости. Рассмотрим регрессионную модель вида (11.2) в случае $p = 1$ и линейной функции

регрессии $f(X)$, т. е.

$$y(x) = \theta_0 + \theta_1 x + \varepsilon(x), \quad (11.24)$$

где θ_0 и θ_1 — неизвестные параметры модели, единственная объясняющая переменная x может быть случайной величиной или неслучайной, $\varepsilon(x)$ и x взаимно некоррелированы, а $E\varepsilon(x) = 0$ и $D\varepsilon(x) = \sigma^2(x)$. В отличие от предыдущей схемы в данном случае не требуется двумерной нормальности анализируемой пары переменных (x, y) . Очевидно, рассматриваемая модель может относиться к любому из типов зависимостей B, C и D , описанных в п. 10.5 (все зависит от конкретизации природы исследуемых переменных).

Определенные соотношения (11.13) и (11.13') — соответственно теоретический и выборочный коэффициенты корреляции — могут быть формально вычислены для любой двумерной системы **линейной** статистической связи между анализируемыми признаками.

Практически все свойства и рекомендации, касающиеся вычисления, использования статистических свойств и интерпретации парного коэффициента корреляции, определенного соотношениями (11.13) и (11.13'), остаются *приблизительно* справедливыми и в рамках более общей схемы (11.24) линейной зависимости между x и y . Так, например, из пяти сформулированных выше свойств коэффициента корреляции первые четыре остаются в силе и в рамках модели (11.24) («не работает» лишь свойство 5: в общем случае из некоррелированности x и y не следует их статистическая независимость).

Однако только в случае совместной нормальной распределенности исследуемых случайных величин x и y коэффициент корреляции r имеет четкий смысл как характеристика степени тесноты связи между ними. В частности, в этом случае соотношение $|r| = 1$ подтверждает чисто функциональную линейную зависимость между исследуемыми величинами, а уравнение $r = 0$ свидетельствует об их полной взаимной независимости. Кроме того, коэффициент корреляции вместе со средними и дисперсиями случайных величин x и y составляет те пять параметров, которые дают исчерпывающие сведения о стохастической зависимости исследуемых величин, так как однозначно определяют их двумерный закон распределения (см. формулу (3.9)). Во всех же остальных случаях (распределения x и y отклоняются от нормального, одна из исследуемых величин не является случайной и т. п.) коэффициент корреляции можно использовать лишь в качестве одной из возможных характеристик степени тесноты связи. При этом, несмотря на то, что в общем случае пока не предложено характеристики линейной связи, которая обладала бы очевидными преимуществами по сравнению с r , его интерпретация часто оказывается весьма ненадеж-

ной. Если же априори *допускается возможность отклонения от линейного вида зависимости, то можно построить примеры, когда, несмотря на $r = 0$, исследуемые переменные оказываются связанными чисто функциональным соотношением* (следовательно, $K_d(x, y) = 1$). Поэтому о величинах, для которых $r = 0$, обычно говорят, что они *некоррелированы*, и только после дополнительного статистического и профессионального анализа (исследование степени отклонения распределения рассматриваемых величин от нормального и т. п.) можно сказать, следует ли отсюда их независимость.

Замечания о необходимости известной осторожности при толковании корреляционной связи никоим образом не обесценивают желательность проверки значимости любого кажущегося соотношения. При этом следует использовать характеристики степени тесноты связи: коэффициента корреляции $\hat{r}$ и корреляционного отношения ρ (см. ниже). Но не всегда знание этих характеристик оказывается достаточным для получения информации о степени тесноты физической связи между исследуемыми переменными и *тем более об их причинной взаимообусловленности*.

11.2.3. Исследование парных нелинейных связей: корреляционное отношение

При отклонениях исследуемой зависимости от линейного вида, как уже отмечалось, коэффициент корреляции r теряет свой смысл как характеристика степени тесноты связи. В этих случаях исследователь должен воспользоваться имеющимися у него двумерными выборочными данными $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ с целью построения оценок для определенной выше в некотором смысле универсальной теоретической характеристики степени тесноты связи — коэффициента детерминации $K_d(y; x)$ (см. (11.8')). Способ построения таких оценок выбирается в зависимости от природы имеющихся у нас выборочных данных и от характера некоторых дополнительных допущений.

Корреляционное отношение. Наиболее привлекательной в этом смысле является ситуация, в которой характер выборочных данных (их количество, «плотность» расположения на плоскости) допускает их группировку по оси объясняющей переменной и возможность подсчета так называемых «частных» средних ординат $\bar{y}_j$ внутри каждого (j -го) интервала группирования. Пусть такое группирование данных произведено. При этом, как обычно, s — число интервалов группирования по оси абсцисс; $\nu_j (j = 1, 2, \dots, s)$ — число выборочных точек, попавших в i -й интервал группирования; $\bar{y}_j = (\sum_{i=1}^{\nu_j} y_{ji}) / \nu_j$ — среднее значение ординат точек,

попавших в j -й интервал группирования. Тогда, как легко понять, выборочным аналогом (оценкой) введенной ранее (см. п. 11.2.1) дисперсии σ_f^2 будет величина

$$s_{\bar{y}(x)}^2 = \frac{1}{n} \sum_{j=1}^s \nu_j (\bar{y}_j - \bar{y})^2, \quad (11.25)$$

где общее среднее $\bar{y} = \frac{1}{n} \sum_{j=1}^s \nu_j \bar{y}_j$. Соответственно получаем оценку для $K_d(y; x)$ в виде

$$\hat{\rho}_{yx}^2 = s_{\bar{y}(x)}^2 / s_y^2, \quad (11.26)$$

где выборочная дисперсия s_y^2 индивидуальных результатов наблюдения y_{ji} около общего среднего $\bar{y}$ вычисляется по формуле

$$s_y^2 = \frac{1}{n} \sum_{j=1}^s \sum_{i=1}^{\nu_j} (y_{ji} - \bar{y})^2.$$

Величину $\hat{\rho}_{yx}$ принято называть *корреляционным отношением зависимой переменной y по независимой переменной x* . Его вычисление не обременено никакими дополнительными допущениями относительно общего вида регрессионной зависимости (11.2). Однако в отличие от коэффициента корреляции *корреляционное отношение несимметрично по отношению к исследуемым переменным*, т. е., вообще говоря, $\rho_{yx} \neq \rho_{xy}$. Кроме того, корреляционное отношение, по определению, является величиной неотрицательной¹, так как под ним подразумевается результат извлечения арифметического значения корня квадратного из ρ^2 .

В остальном свойства корреляционного отношения во многом похожи на свойства коэффициента корреляции. Из (11.3) и (11.8), в частности, немедленно следует, что подобно коэффициенту корреляции корреляционное отношение не может быть больше единицы.

Из $|\rho| = 1$ следует наличие однозначной функциональной связи между y и x , и, наоборот, однозначная функциональная связь между y и x свидетельствует о том, что $|\rho| = 1$. Далее, отсутствие корреляционной связи между y и x означает, что условные средние $\bar{y}_j$ сохраняют постоянное значение, равное общему среднему $\bar{y}$, а потому $\rho_{yx} = 0$. Наоборот, если $\rho_{yx} = 0$, то $\bar{y}_j = \bar{y}$, и, следовательно, частные средние $\bar{y}_j$ не зависят от x , т. е. соответствующая линия регрессии параллельна горизонтальной оси.

¹ Иногда, в частности при *монотонном* характере функции регрессии $f(x)$, корреляционному отношению приписывают знак, совпадающий со знаком первой производной этой функции.

Отметим, что между ρ_{yx} и ρ_{xy} нет какой-либо простой зависимости. Некоррелированность y с x (т.е. равенство нулю величины ρ_{yx}) не влечет за собой непосредственно некоррелированности x с y . Возможны ситуации, в которых один из этих показателей принимает нулевое значение, в то время как другой равен единице. Допустим, например, что $y = x^2$ и x принимает значения: $-1, 0$ и $+1$ с вероятностями $1/3$ каждое. В этом случае $\rho_{yx} = 1$, $\rho_{xy} = 0$ (в силу симметрии параболы относительно оси y и симметричности распределения x).

Можно показать, что корреляционное отношение ρ не может быть меньше абсолютной величины коэффициента корреляции r , характеризующего зависимость между теми же переменными. В случае линейной зависимости эти две характеристики связи совпадают. Это позволяет использовать величину разности $\hat{\rho}_{yx}^2 - \hat{r}^2$ в качестве меры отклонения регрессионной зависимости от линейного вида (см. соотношение (10.25) в п. 10.7).

И наконец, все замечания относительно смысловой интерпретации коэффициента корреляции r (в частности, о логическом соотношении понятий «корреляционная зависимость, связь между переменными, их причинная взаимообусловленность») остаются в силе и для корреляционного отношения.

Проверка гипотезы об отсутствии корреляционной связи. Какую величину корреляционного отношения можно признать статистически значимо отличающейся от нуля, т.е. достаточной для статистически обоснованного вывода о наличии корреляционной связи между исследуемыми переменными? Ведь так же, как и в случае прямолинейного типа зависимости, принципиально возможны ситуации, когда отклонение от нуля полученной величины корреляционного отношения $\hat{\rho}$ является статистически незначимым, т.е. обусловленным лишь неизбежными случайными колебаниями выборки. Для построения соответствующего критерия воспользуемся фактом приближенной $F(s-1, n-s)$ -распределенности случайной величины

$$\hat{F}(0) = \frac{\hat{\rho}_{yx}^2}{1 - \hat{\rho}_{yx}^2} \cdot \frac{n-s}{s-1}, \quad (11.27)$$

справедливым в предположении, что $K_d(y; x) = 0$ (или, что то же, $\rho_{yx} = 0$) и что условные распределения результирующей переменной $y(x)$ при любом фиксированном x описываются нормальным законом с постоянной дисперсией σ^2 .

Поэтому если окажется, что

$$\frac{\hat{\rho}_{yx}^2}{1 - \hat{\rho}_{yx}^2} \cdot \frac{n - s}{s - 1} > v_{\alpha}^2(s - 1, n - s),$$

то гипотеза $H: \rho_{yx} = 0$ об отсутствии корреляционной связи между y и x отвергается с вероятностью ошибки α (здесь, как и ранее, $v_{\alpha}^2(s - 1, n - s)$ — $100\alpha\%$ -ная точка F -распределения с числом степеней свободы числителя $s - 1$ и знаменателя $n - s$ находится из табл. приложения 1). При выполнении обратного неравенства значение корреляционного отношения $\hat{\rho}_{yx}$ признается статистически незначимым, т.е. делается вывод, что гипотеза об отсутствии корреляционной связи между y и x не противоречит наблюдениям.

Доверительные интервалы для истинного значения корреляционного отношения ρ_{yx} можно построить, опираясь на тот факт, что статистика

$$\gamma^{(n)} = \frac{(n - s)\hat{\rho}_{yx}^2}{(s - 1)(1 - \hat{\rho}_{yx}^2)} \cdot \frac{s - 1}{s - 1 + n\hat{\rho}_{yx}^2}$$

приблизительно описывается F -распределением с числом степеней свободы числителя

$$\nu_1^* = \frac{(s - 1 + n\hat{\rho}_{yx}^2)^2}{s - 1 + 2n\hat{\rho}_{yx}^2} \quad (11.28)$$

и числом степеней свободы знаменателя $\nu_2 = n - s$ (см. [Айвазян С. А., Енюков И. С., Мешалкин Л. Д., 1985, п.1.1.4]).

Таким образом, получаем следующее правило построения приближенных доверительных интервалов для истинного значения корреляционного отношения ρ_{yx} :

1) пользуясь формулой (11.26), вычисляем точечную оценку $\hat{\rho}_{yx}^2$ для истинного значения корреляционного отношения ρ_{yx} ;

2) по формуле (11.28) подсчитываем вспомогательное число степеней свободы ν_1^* числителя для аппроксимирующего F -распределения;

3) задавшись уровнем доверия $P = 1 - 2\alpha$, с помощью табл. приложения 1 находим $100(1 - \alpha)\%$ -ную точку $v_{1-\alpha}^2(\nu_1^*, n - s)$ и $100\alpha\%$ -ную точку $v_{\alpha}^2(\nu_1^*, n - s)$ F -распределения с числом степеней свободы числителя ν_1^* и знаменателя $n - s$;

4) утверждаем, что приблизительно с вероятностью $P = 1 - 2\alpha$ истинное значение корреляционного отношения ρ_{yx} удовлетворяет неравенствам

$$\frac{(n - s)\hat{\rho}_{yx}^2}{n(1 - \hat{\rho}_{yx}^2)v_{\alpha}^2} - \frac{s - 1}{n} < \rho_{yx} < \frac{(n - s)\hat{\rho}_{yx}^2}{n(1 - \hat{\rho}_{yx}^2)v_{1-\alpha}^2} - \frac{s - 1}{n}. \quad (11.29)$$

Проиллюстрируем работоспособность описанного метода на следующем примере. Пусть в результате обработки 132 экспериментальных точек $(x_i, y_i) (i = 1, 2, \dots, 132)$ получено выборочное значение корреляционного отношения $\hat{\rho} = 0,60$. При этом мы воспользовались разбиением диапазона изменения независимой переменной на $s = 12$ равных интервалов группирования. Соответственно получаем в качестве вспомогательного числа степеней свободы числителя величину $\nu_1^* = \frac{(12-1+132 \cdot 0,36)^2}{12-1+2 \cdot 132 \cdot 0,36} \approx 27$ (частное округляем до целого числа). Задавшись доверительной вероятностью $P = 0,90$, из табл. приложения 1 находим (полагая $\alpha = 0,05$):

$$v_{0,05}^2(27, 120) = 1,58;$$

$$v_{0,95}^2(27, 120) = \frac{1}{v_{0,05}^2(120, 27)} = \frac{1}{1,73} \approx 0,58.$$

И наконец, в соответствии с формулой (11.29) находим левый ($\hat{\rho}_{\min}^2$) и правый ($\hat{\rho}_{\max}^2$) концы доверительного интервала для истинного значения ρ_{yx} :

$$\hat{\rho}_{\min}^2 = \frac{120 \cdot 0,36}{132 \cdot 0,64 \cdot 1,58} - \frac{11}{132} = 0,24;$$

$$\hat{\rho}_{\max}^2 = \frac{120 \cdot 0,36}{132 \cdot 0,64 \cdot 1,58} - \frac{11}{132} = 0,87.$$

Таким образом, при точечной оценке $\hat{\rho}_{yx} = 0,6$ истинное значение заключено в пределах от $\sqrt{0,24}$ до $\sqrt{0,87}$ с вероятностью, приблизительно равной 0,9, т. е. $0,49 < \rho_{yx} < 0,93$.

В этом примере хорошо видна *существенная несимметричность концов интервальной оценки* относительно точечной оценки (правый конец интервальной оценки отстоит от точечной оценки на 0,33, в то время как левый конец — всего лишь на 0,11).

Для значений точечных оценок $\hat{\rho}^2$, близких к нулю или к единице, левый или правый конец интервальной оценки может терять содержательный смысл, выходя за пределы отрезка $[0, 1]$. В этом случае в качестве левого или правого конца интервальной оценки следует брать соответствующее граничное значение — нуль или единицу (причина подобных нежелательных ситуаций — в аппроксимационном подходе к решению данной задачи). Однако описанный прием все-таки следует признать гораздо более точным, чем применяемый иногда метод построения интервальных оценок для $\rho_{\eta\xi}$, необоснованно использующий приближительную $(\rho, \frac{1-\rho^2}{\sqrt{n}})$ -нормальность статистики $\hat{\rho}_{\eta\xi}$.

11.2.4. Исследование линейной зависимости y от нескольких объясняющих переменных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$: множественный и частные коэффициенты корреляции

Аналогично той схеме, по которой излагался п. 11.2.2, разобьем рассмотрение данной проблемы на два случая: 1) «поведение» вектора переменных $(x^{(1)}, x^{(2)}, \dots, x^{(p)}; y)$ описывается $(p+1)$ -мерным нормальным законом (см. формулу (3.10) в п. 3.1.1); 2) результирующий показатель y связан с объясняющими переменными $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ линейной регрессионной зависимостью типа (11.2).

1) **Случай многомерной нормальности векторного признака** $(x^{(1)}, x^{(2)}, \dots, x^{(p)}; y)$. Свойства многомерного нормального закона достаточно хорошо исследованы (см., например, [Андерсон Т.]). В частности, имеют место результаты, аналогичные (11.10)–(11.14), которые были рассмотрены в рамках двумерной схемы в п. 11.2.2. А именно доказано, что функция регрессии y по X имеет *линейный вид*:

$$f(X) = E(y | X) = \theta_0 + \theta_1 x^{(1)} + \dots + \theta_p x^{(p)}, \quad (11.30)$$

где коэффициенты θ_j ($j = 0, 1, \dots, p$) явным образом выражаются в терминах компонент вектора средних значений M_1 и элементов ковариационной матрицы Σ анализируемого многомерного нормального распределения, а условная дисперсия y (при условии, что объясняющие переменные зафиксированы на уровне X) *не зависит от X* и тоже выражается в явном виде через элементы ковариационной матрицы Σ .

Однако в анализе *множественных* корреляционных связей (так называют статистические связи между более чем двумя переменными в отличие от парных связей, рассмотренных выше) есть своя специфика и возникают принципиально новые проблемы. Эта специфика связана в первую очередь с необходимостью уметь измерять степень тесноты связи между результирующей переменной y и *множеством объясняющих переменных* $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, а также с возникающими трудностями в интерпретации парных коэффициентов корреляции между y и $x^{(j)}$, обусловленными возможным опосредованным влиянием на эту парную связь других (явно не учтенных в вычислении $r(y; x^{(j)})$) объясняющих переменных $x^{(i)}$ ($i \neq j$). Последнее обстоятельство, в частности, делает необходимым введение таких измерителей статистической связи, которые были бы «очищены» от опосредованного влияния других переменных, давали бы оценку степени тесноты интересующей нас связи между переменными y и $x^{(j)}$ (или $x^{(i)}$ и $x^{(j)}$) при условии, что значения остальных переменных зафиксированы

на некотором постоянном уровне. В этом случае говорят о статистическом анализе *частных* (или «очищенных») связей и используют соответственно *частные* («очищенные») коэффициенты корреляции или другие корреляционные характеристики.

Частные коэффициенты корреляции и их выборочные значения. Поставим в соответствие каждой из ранее введенных парных характеристик статистической связи между переменными $x^{(i)}$ и $x^{(j)}$ ($i, j = 0, 1, \dots, p$; $x^{(0)} \equiv y$) *частную* («очищенную») характеристику, определяемую по той же формуле, но только для *условного* распределения (см. п. 2.5, (2.13)) $\varphi(x^{(i)}, x^{(j)} | X^{(i,j)} = x)$. Здесь φ — это функция плотности вероятности переменных $x^{(i)}$ и $x^{(j)}$; $X^{(i,j)}$ — множество переменных, дополняющих пару $(x^{(i)}, x^{(j)})$ до *полного* набора рассматриваемых (наблюдаемых) переменных $X = (x^{(0)}, x^{(1)}, \dots, x^{(p)})$, а x — $(p-1)$ -мерный вектор, определяющий заданные уровни, на которых фиксируются значения «мешающих» переменных $X^{(i,j)}$. Есть два взаимосвязанных обстоятельства, которые препятствуют широкому практическому использованию частных характеристик статистической связи в *общем* (т. е. негауссовском) *случае*:

- частные характеристики статистической связи, вообще говоря, зависят от заданных уровней x мешающих переменных (как их выбирать в каждом конкретном случае?);
- для подсчета *выборочных* значений частных характеристик статистической связи необходимо иметь выборку *специальной структуры*, обеспечивающей наличие хотя бы нескольких наблюдений при каждом из заданного ряда фиксированных значений x мешающих переменных.

Однако можно показать, что если исследуемые случайные переменные $(x^{(0)}, x^{(1)}, \dots, x^{(p)})$ подчиняются *многомерному нормальному закону*, то указанные неудобства автоматически исчезают, так как в этом случае частные коэффициенты корреляции не зависят от уровней мешающих переменных x , определяющих условие в соответствующем условном распределении. В частности, имеет место следующая формула (при условии невырожденности $(p+1)$ -мерного нормального закона):

$$r_{ij, X^{(i,j)}} = \frac{-R_{ij}}{(R_{ii}R_{jj})^{\frac{1}{2}}}, \quad (11.31)$$

где $r_{ij, X^{(i,j)}}$ — частный коэффициент корреляции между переменными $x^{(i)}$ и $x^{(j)}$ при фиксированных значениях всех остальных переменных $X^{(i,j)}$, а R_{kl} — алгебраическое дополнение (см. Приложение 2) для элемента

r_{kl} в определителе корреляционной матрицы $\mathbf{R}$ анализируемых признаков $x^{(0)} \equiv y, x^{(1)}, x^{(2)}, \dots, x^{(p)}$, т. е. в определителе

$$\det \mathbf{R} = \begin{vmatrix} 1 & r_{01} & r_{02} & \dots & r_{0p} \\ r_{10} & 1 & r_{12} & \dots & r_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ r_{p0} & r_{p1} & r_{p2} & \dots & 1 \end{vmatrix}, \quad \text{где } r_{ij} = r(x^{(i)}, x^{(j)}).$$

Формула (11.31), примененная к *трехмерному* признаку ($x^{(0)} \equiv y, x^{(1)}, x^{(2)}$), при $i = 0, j = 1$ и $X^{(i,j)} = x^{(2)}$ дает:

$$r_{01 \cdot x^{(2)}} = r_{01(2)} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{(1 - r_{02}^2)(1 - r_{12}^2)}}. \quad (11.32)$$

Последовательно присоединяя к мешающим переменным все новые признаки из рассматриваемого набора, можно получить *рекуррентные* соотношения для подсчета частных коэффициентов корреляции $r_{01(2 \dots k+1)}$ порядка k (т. е. при исключении опосредованного влияния k мешающих переменных) по частным коэффициентам корреляции порядка $k - 1$ ($k = 1, 2, \dots, p - 1$):

$$r_{01(2,3,\dots,k+1)} = \frac{r_{01(2,3,\dots,k+1)} - r_{0k+1(2,\dots,k)} \cdot r_{1k+1(2,\dots,k)}}{\sqrt{(1 - r_{0k+1(2,\dots,k)}^2)(1 - r_{1k+1(2,\dots,k)}^2)}}. \quad (11.32')$$

Выборочные (эмпирические) значения частных коэффициентов корреляции вычисляются по тем же формулам (11.31) – (11.32') с заменой теоретических значений парных коэффициентов корреляции r_{ij} их выборочными аналогами $\hat{r}_{ij}$ (см. формулу (11.13')).

Если исследователь имеет дело лишь с тремя-четырьмя переменными ($p = 2, 3$), то удобно пользоваться рекуррентными соотношениями (11.32'). При больших размерностях анализируемого многомерного признака удобнее опираться на формулу (11.31), использующую расчет соответствующих определителей.

Статистические свойства выборочных частных коэффициентов корреляции (проверка на статистическую значимость их отличия от нуля, доверительные интервалы). При исследовании статистических свойств выборочного частного коэффициента корреляции порядка k (т. е. при исключении опосредованного влияния k мешающих переменных) следует воспользоваться тем, что он распределен точно так же, как и обычный (парный) выборочный коэффициент корреляции между теми же переменными с единственной поправкой: объем выборки надо уменьшить на

k единиц, т. е. полагать его равным $n - k$, а не n . Поэтому при проверке статистически значимого отличия от нуля выборочного частного коэффициента корреляции и при построении для него доверительных интервалов следует пользоваться рекомендациями п. 11.2.2 для парного коэффициента корреляции с заменой n на $n - k$.

Рассмотрим некоторые конкретные числовые примеры, демонстрирующие возможный характер искажающего опосредованного влияния «третьих факторов» на корреляцию между двумя анализируемыми переменными.

Пример 11.3. По итогам года 37 однородных предприятий легкой промышленности были зарегистрированы следующие показатели их работы: $x^{(0)} \equiv y$ — среднемесячная характеристика качества ткани (в баллах); $x^{(1)}$ — среднемесячное количество профилактических наладок автоматической линии; $x^{(2)}$ — среднемесячное число обрывов нити.

По матрице исходных данных $(x_i^{(0)}, x_i^{(1)}, x_i^{(2)})_{i=1,37}$ были подсчитаны (с помощью (11.13')) выборочные *парные* коэффициенты корреляции $\hat{r}_{ij}(i, j = 0, 1, 2)$: $\hat{r}_{01} = 0,105$; $\hat{r}_{02} = 0,024$; $\hat{r}_{12} = 0,996$.

Проверка «на статистическую значимость», проведенная в соответствии с рекомендациями п. 11.2.2, свидетельствует об отсутствии статистически значимой парной корреляционной связи между качеством ткани, с одной стороны, и числом профилактических наладок и обрывов нити — с другой, что не согласуется с профессиональными представлениями технолога.

Однако расчет *частных* коэффициентов корреляции по формуле (11.32) дает значения $\hat{r}_{01(2)} = 0,907$; $\hat{r}_{02(1)} = -0,906$, которые вполне соответствуют нашим представлениям о естественном характере связей между изучаемыми показателями.

Доверительные интервалы для *истинных* значений $r_{01(2)}$ и $r_{02(1)}$ (в соответствии с рекомендациями п. 11.2.2) найдем с использованием z -преобразования Фишера для доверительной вероятности $P = 1 - \alpha$. Тогда

$$z_{1,2} = \frac{1}{2} \ln \frac{1 + \hat{r}}{1 - \hat{r}} \mp \frac{u_{\alpha/2}}{\sqrt{(n-1)-3}} - \frac{\hat{r}}{2[(n-1)-1]},$$

где u_q — q -квантиль стандартного нормального распределения (см. табл. Приложения 1).

В нашем примере $n = 37$, $\alpha = 0,05$. Подставляя поочередно в эту формулу значения $\hat{r}_{01(2)} = 0,907$ и $\hat{r}_{02(1)} = -0,906$ и пользуясь табл. П1.7 значений $\hat{z} = \operatorname{arcth} \hat{r} = \frac{1}{2} \ln \frac{1+\hat{r}}{1-\hat{r}}$, найдем для

$$\hat{r} = \hat{r}_{01(2)} = 0,907 :$$

$$z_1 = 1,16$$

и

$$z_2 = 1,83,$$

а для

$$\hat{r} = \hat{r}_{02(1)} = -0,906; z_1 = -1,83$$

и

$$z_2 = -1,16,$$

откуда, вновь воспользовавшись табл., окончательно получим:

$$0,820 < r_{01(2)} < 0,950;$$

$$-0,950 < r_{02(1)} < -0,820.$$

Пример 11.4. С целью исследования влияния погодных условий на урожайность кормовых трав Хукер (Journ. Roy. Stat. Soc., 1907, v. 65, p. 1) рассмотрел данные Министерства земледелия Англии за 20 лет, характеризующие урожайность $x^{(0)}$ (в ц/акр), весеннее количество осадков $x^{(1)}$ (в дюймах) и накопленную за весну сумму «активных» (т.е. выше $+5,5^\circ\text{C}$) температур $x^{(2)}$ (в градусах по Фаренгейту) однородной в метеорологическом отношении области Англии, включающей в себя группу восточных графств. По выборке $(x_i^{(0)}, x_i^{(1)}, x_i^{(2)})_{i=1,20}$ были подсчитаны основные статистические характеристики изучаемой трехмерной случайной величины:

$$\begin{aligned} \hat{m}_1^{(0)} &= 28,02; & \hat{m}_1^{(1)} &= 4,91; & \hat{m}_1^{(2)} &= 594,0; \\ s_0^2 &= 19,54; & s_1^2 &= 1,21; & s_2^2 &= 7225; \\ \hat{r}_{01} &= 0,80; & \hat{r}_{02} &= -0,40; & \hat{r}_{12} &= -0,56. \end{aligned}$$

Действительно ли высокая температура в период созревания трав отрицательно влияет на их урожайность (ведь $\hat{r}_{02} = -0,40$) или здесь сказывается опосредованное влияние «мешающего» фактора — количества осадков $x^{(1)}$?

Вычисление частных коэффициентов корреляции по рекуррентной формуле (11.32) дает:

$$\hat{r}_{01(2)} = 0,759; \hat{r}_{02(1)} = -0,097; \hat{r}_{12(0)} = -0,436.$$

Как видим, если исключить одновременное влияние количества осадков $x^{(1)}$ на урожайность (с ростом $x^{(1)}$ она повышается) и на сумму активных температур (с ростом $x^{(1)}$ она понижается), то мы уже не обнаружим отрицательной корреляции между температурой и урожайностью ($\hat{r}_{02(1)} = 0,097$, в то время как $\hat{r}_{02} = -0,40$).

Построение доверительных интервалов для $r_{01(2)}$ и $r_{02(1)}$ (с уровнем доверия $P = 0,95$) с использованием z -преобразования Фишера дает в данном случае:

$$0,448 < r_{01(2)} < 0,890; \quad -0,419 < r_{02(1)} < 0,525.$$

Последнее неравенство свидетельствует о том, что у нас нет оснований считать положительную очищенную корреляционную связь между урожайностью и температурой ($r_{02(1)} = 0,097$) статистически значимой, т. к. нуль находится внутри доверительного интервала.

Обратимся теперь к задаче измерения с.т.с.с. между результирующим показателем y и множеством объясняющих переменных $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^T$ в условиях многомерной нормальной совокупности. В общем случае эта задача решается с помощью коэффициента детерминации $K_d(y, X)$ (см. (11.8) и (11.8')), который по построению обладает следующими свойствами:

а) $0 \leq K_d(y; X) \leq 1$;

б) минимальное значение коэффициента детерминации ($K_d(y; X) = 0$) соответствует случаю полного отсутствия корреляционной связи между y и $(x^{(1)}, \dots, x^{(p)})$, так как это может быть только при $\sigma_f^2 = Df(\xi) = 0$, т. е. при независимости значений функции регрессии f от величины ее аргументов $X(f(X) = \text{const})$; это соответствует ситуации, когда усредненная дисперсия «регрессионных остатков» в точности равна общей вариации результирующего показателя;

в) максимальное значение коэффициента детерминации ($K_d(y; X) = 1$) соответствует полному отсутствию варьирования «регрессионных остатков» ($E\varepsilon^2 = 0$), что означает наличие чисто функциональной связи между y и $(x^{(1)}, \dots, x^{(p)})$: $y = f(x^{(1)}, \dots, x^{(p)})$. Следовательно, в этом случае мы имеем возможность *точно (детерминированно)* восстанавливать условные значения $y(X) = \{y \mid X\}$ по значениям предикторных переменных X , и соответственно общая вариация результирующего показателя y *полностью объясняется* контролируемой вариацией функции регрессии.

Из (11.8') следует, что вычисление выборочного значения коэффициента детерминации предусматривает проведение предварительных расчетов по статистическому оцениванию неизвестной функции регрессии $f(X)$, что противоречит хронологии исследований, описанной в п. 10.6. Сейчас мы увидим, что свойства многомерных *нормальных* совокупностей автоматически устраняют это неудобство, позволяя вычислять значение $K_d(y; X)$ до проведения регрессионного анализа.

Множественный коэффициент корреляции $R_{y,X}$ используется в качестве измерителя с.т.с.с. между результирующим показателем y и

набором объясняющих переменных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ в моделях линейной регрессии. Он определяется как обычный парный коэффициент корреляции между y и линейной функцией регрессии y по X , т. е.

$$R_{y.X} = r(y; f(X)), \quad (11.33)$$

где

$$f(X) = E(y | X) = \theta_0 + \theta_1 x^{(1)} + \dots + \theta_p x^{(p)}$$

(можно показать, что следующее определение $R_{y.X}$ является эквивалентным определению (11.33): $R_{y.X}$ — это парный коэффициент корреляции между y и такой линейной комбинацией $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, для которой значение этого парного коэффициента корреляции достигает своего максимума).

Можно показать, что при статистической обработке выборок, извлеченных из нормальных генеральных совокупностей, множественный коэффициент корреляции $R_{y.X}$ и его выборочное значение $\hat{R}_{y.X}$ обладают рядом удобных свойств (приведенные ниже формулы и свойства теоретического множественного коэффициента корреляции $R_{y.X}$ автоматически переносятся на выборочный $\hat{R}_{y.X}$ заменой участвующих в них теоретических характеристик соответствующими выборочными значениями).

1. *Вычисление $R_{y.X}$ по матрице парных коэффициентов корреляции.* Обозначая, как и прежде, $(p+1)(p+1)$ -корреляционную матрицу $(r_{ij})_{i,j=0,1,\dots,p}$ через $\mathbf{R}$, а алгебраическое дополнение элемента r_{kl} в ее определителе через R_{kl} , имеем

$$R_{y.X}^2 = 1 - \frac{\det \mathbf{R}}{R_{00}}. \quad (11.34)$$

2. *Вычисление $R_{y.X}$ по частным коэффициентам корреляции*

$$R_{y.X}^2 = 1 - (1 - r_{01}^2)(1 - r_{02(1)}^2)(1 - r_{03(12)}^2) \dots (1 - r_{0p(12\dots p-1)}^2). \quad (11.35)$$

3. *Множественный коэффициент корреляции мажорирует любой парный или частный коэффициент корреляции, характеризующий статистическую связь результирующего показателя, т. е.*

$$R_{y.X} \geq |r_{0j(I_j)}|,$$

где $j = 1, 2, \dots, p$, а I_j — любое подмножество множества индексов $I_0 = \{1, 2, \dots, p\}$, не содержащее индекса j (это соотношение следует из (11.35)). Напоминаем, что $x^{(0)} \equiv y$.

4. Присоединение каждой новой предсказывающей переменной не может уменьшить величины R (независимо от порядка присоединения), т. е.

$$R_{y,x(1)} \leq R_{y,(x^{(1)},x^{(2)})} \leq R_{y,(x^{(1)},x^{(2)},x^{(3)})} \leq \dots \leq R_{y,(x^{(1)},x^{(2)},\dots,x^{(p)})}. \quad (11.37)$$

5. Условная дисперсия результирующего показателя $D(y | X)$ не зависит от условия (т. е. от значения X) и связана со значением множественного коэффициента корреляции $R_{y,X}$ соотношением

$$D(y | X) = (1 - R_{y,X}^2) Dy. \quad (11.36)$$

Последний результат позволяет связать между собой две характеристики степени тесноты статистической связи — коэффициент детерминации $K_d(y; X)$ и множественный коэффициент корреляции $R_{y,X}$. Действительно, учитывая тот факт, что $D(y | X) = D(\varepsilon | X)$ (см. (11.5)), а значит, и условная дисперсия остатков не зависит от X , получаем, что безусловная дисперсия $D\varepsilon$ (как результат усреднения по X значений условных дисперсий $D(\varepsilon | X)$, см. (11.9)) равна $D(\varepsilon | X)$. Но тогда мы можем подставить в (11.36) $D\varepsilon$ вместо $D(y | X)$ и получим

$$D\varepsilon = (1 - R_{y,X}^2) Dy, \quad (11.36')$$

откуда, учитывая формулу (11.8),

$$K_d(y; X) = R_{y,X}^2. \quad (11.37)$$

Это означает, что в рамках статистического анализа многомерной нормальной совокупности понятие введенного соотношением (11.8) коэффициента детерминации $K_d(y; X)$ совпадает с квадратом определенного в (11.33) множественного коэффициента корреляции $R_{y,X}^2$ и что коэффициент детерминации $K_d(y; X)$ может быть вычислен в данном случае до проведения регрессионного анализа (т. е. до оценки функции регрессии $f(X)$) с помощью формул (11.34), (11.35).

Для проверки гипотезы $H_0: R_{y,X} = 0$, т. е. для выяснения вопроса, можно ли считать выборочное значение множественного коэффициента корреляции $\hat{R}_{y,X}$ статистически значимо отличающимся от нуля, пользуются фактом $F(p, n - p - 1)$ -распределенности случайной величины

$$F(\hat{R}) = \frac{\hat{R}_{y,X}^2}{1 - \hat{R}_{y,X}^2} \cdot \frac{n - p - 1}{p},$$

справедливым в рамках рассматриваемой многомерной нормальной совокупности при условии, что истинное значение множественного коэффициента корреляции $R_{y,X}$ равно нулю. Если окажется, что $\hat{F}(\hat{R}) >$

$v_{\alpha}^2(p, n-p-1)$, то гипотеза об отсутствии множественной корреляционной связи между y и $(x^{(1)}, x^{(2)}, \dots, x^{(p)})$ отвергается при уровне значимости критерия, равном α (здесь, как и ранее, $v_{\alpha}^2(p, n-p-1)$ — $100\alpha\%$ -ная точка F -распределения с числом степеней свободы числителя p и знаменателя $n-p-1$ находится из табл. Приложения 1).

Знание следующих статистических свойств оценок $\hat{R}_{y.X}$, определяемых по формулам (11.34), (11.35) с заменой участвующих в них парных и частных коэффициентов корреляции их выборочными аналогами, может оказаться полезным при проведении корреляционного и регрессионного анализов:

$$\bullet \quad E\hat{R}_{y.X}^2 \approx R_{y.X}^2 + \frac{p}{n-1}(1 - R_{y.X}^2), \quad (11.38)$$

что означает наличие положительного смещения (асимптотически устранимого) у оценки $\hat{R}_{y.X}^2$ коэффициента детерминации $R_{y.X}^2$;

$$\bullet \quad D\hat{R}_{y.X}^2 \approx \frac{2p(n-p-1)}{(n-1)(n^2-1)}(1 - R_{y.X}^2)^2 \quad (11.39)$$

(дисперсия $D\hat{R}_{y.X}^2$ характеризует *точность оценивания* коэффициента детерминации $R_{y.X}^2$ с помощью $\hat{R}_{y.X}^2$ и будет использована в регрессионном анализе при определении числа объясняющих переменных, которые следует включить в линейную регрессионную модель);

• подправленная на несмещенность оценка $\hat{R}_{y.X}^{*2}$ коэффициента детерминации $R_{y.X}^2$ имеет вид

$$\hat{R}_{y.X}^{*2} \approx 1 - (1 - \hat{R}_{y.X}^2) \frac{n-1}{n-p-1}. \quad (11.40)$$

Из последней формулы видно, что «подправленная» оценка $\hat{R}_{y.X}^{*2}$ всегда меньше смещенной оценки $\hat{R}_{y.X}^2$.

Отметим, что при малых истинных значениях $R_{y.X}^2$ и при «не слишком малых» величинах отношения p/n подправленные оценки, подсчитанные по формуле (11.40), могут принимать *отрицательные* значения. Можно устранить абсурдность отрицательных значений оценки, используя в качестве «еще раз подправленной» оценки величину

$$\hat{R}_{y.X}^{**2} = \max(\hat{R}_{y.X}^{*2}, 0)$$

(правда, $\hat{R}_{y.X}^{**2}$ уже не будет несмещенной оценкой).

Вернемся к ранее рассмотренным примерам и оценим в них степень тесноты множественной связи между результирующим показателем, с одной стороны, и набором объясняющих переменных — с другой. Будем

пользоваться рекомендациями (а именно формулами (11.34), (11.35)), правомерность которых, напомним, строго обоснована лишь в рамках многомерной нормальной генеральной совокупности.

Пример 11.3 (продолжение). Оценка $\hat{R}_{y,(x^{(1)}x^{(2)})}$ коэффициента множественной корреляции между характеристикой качества ткани y и совокупностью двух факторов: количеством профилактических наладок $x^{(1)}$ и числом обрывов нити $x^{(2)}$, подсчитанная с помощью формулы (11.35), дает:

$$\begin{aligned}\hat{R}_{y,(x^{(1)}x^{(2)})}^2 &= 1 - (1 - \hat{r}_{01}^2)(1 - \hat{r}_{02(1)}^2) \\ &= 1 - [1 - (0,105)^2][1 - (0,906)^2] \\ &= 1 - 0,989 \cdot 0,179 = 1 - 0,177 = 0,823.\end{aligned}$$

Отсюда $\hat{R}_{y,(x^{(1)}x^{(2)})} = \sqrt{0,823} = 0,9072$.

Пример 11.4 (продолжение). Оценка $\hat{R}_{y,(x^{(1)}x^{(2)})}$ коэффициента множественной корреляции между урожайностью кормовых трав ($y \equiv x^{(0)}$) и природными факторами — весенним количеством осадков ($x^{(1)}$) и накопленной суммой «активных» температур ($x^{(2)}$), подсчитанная по формуле (11.35), дает:

$$\begin{aligned}\hat{R}_{y,(x^{(1)}x^{(2)})}^2 &= 1 - (1 - \hat{r}_{01}^2)(1 - \hat{r}_{02(1)}^2) \\ &= 1 - [1 - (0,80)^2][1 - (1,097)^2] \\ &= 1 - 0,36 \cdot 0,99 = 0,6436.\end{aligned}$$

Отсюда $\hat{R}_{y,(x^{(1)}x^{(2)})} = \sqrt{0,6436} = 0,802$.

2) Исследование линейных множественных связей (общий случай). Речь идет о корреляционном анализе линейной модели вида

$$y = \theta_0 + \theta_1 x^{(1)} + \dots + \theta_p x^{(p)} + \varepsilon(X), \quad (11.41)$$

где $\Theta = (\theta_0, \theta_1, \dots, \theta_p)^\top$ — неизвестные параметры модели, объясняющие переменные, могут быть как случайными (схемы зависимостей типов C и D), так и неслучайными (схема зависимости типа B , см. п. 10.5), $\varepsilon(X)$ и $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})$, взаимно некоррелированы, а $E\varepsilon(X) \equiv 0$, $D\varepsilon(X) = \sigma^2(X)$. В отличие от предыдущего случая в модели (11.41) не требуется совместной многомерной нормальности переменных $(y, x^{(1)}, x^{(2)}, \dots, x^{(p)})$.

В статистической практике свойства и рекомендации, справедливые в условиях многомерной нормальной совокупности (относящиеся к частным и множественным коэффициентам корреляции), обычно распространяют

и на общий случай (11.41). И, как правило, частные коэффициенты корреляции, определенные соотношениями (11.31)–(11.32'), являются удовлетворительными измерителями очищенной линейной связи между $x^{(i)}$ и $x^{(j)}$ при фиксированных значениях остальных переменных $X^{(i,j)}$ и в случае, когда распределение анализируемых показателей $(x^{(0)}, x^{(1)}, \dots, x^{(p)})$ отличается от нормального. При этом их можно интерпретировать как показатели тесноты очищенной связи, усредненные по всевозможным значениям фиксируемых на определенных уровнях «мешающих» переменных.

Точно так же множественный коэффициент корреляции $R_{y,X}$, определенный формулами (11.34)–(11.35), может быть формально вычислен для любой многомерной системы наблюдений. И в случае линейной зависимости (11.41) результаты вычисления коэффициента детерминации, полученные по этим формулам, относительно слабо отличаются от результата «прямого счета» по формуле (11.8'). Однако даже при незначительных отклонениях типа функции регрессии $f(X)$ от линейной доверять значению $\hat{R}_{y,X}^2$ как аппроксимации для величины коэффициента детерминации $K_d(y; X)$ не рекомендуется. В этом случае следует пользоваться непосредственно формулой (11.8'), требующей предварительной оценки функции регрессии $f(X)$ в точках $X_1, X_2, \dots, X_n$. Возникающие при этом неудобства (особенно в условиях отсутствия априорной информации об общем виде функции регрессии $f(X)$) преодолевают, в частности, следующим образом:

а) разбивают область возможных значений объясняющих переменных на многомерные аналоги интервалов группирования — гиперпараллелепипеды группирования $\Delta_1, \Delta_2, \dots, \Delta_s$;

б) подсчитывают условные средние $\bar{y}_j$ результирующего показателя по наблюдениям, попавшим в j -й гиперпараллелепипед группирования ($j = 1, 2, \dots, s$);

в) по наблюдениям, попавшим в Δ_j , оценивают условную дисперсию результирующего показателя s_j^2 по отклонениям этих наблюдений от своего условного среднего $\bar{y}_j$ ($j = 1, 2, \dots, s$);

г) с помощью взвешенного усреднения условных дисперсий s_j^2 ($j = 1, 2, \dots, s$) оценивают безусловную дисперсию остатков $D\varepsilon$, участвующую в формуле (11.8');

д) оценив общую дисперсию $s_y^2 = (\Sigma(y_i - \bar{y})^2)/n$ результирующего показателя, подставляют оценку $D\varepsilon$ и s_y^2 в формулу (11.8') и вычисляют $\hat{K}_d(y, X)$.

По существу эта процедура является многомерным аналогом процедуры вычисления парного корреляционного отношения.

11.3. Корреляционный анализ порядковых (ординарных) переменных: ранговая корреляция

Напомним (см. п. 2.3), что *порядковая (ординальная)* переменная позволяет *упорядочивать* статистически обследованные объекты по степени проявления в них анализируемого свойства. Исследователь обращается к порядковым переменным в ситуациях, когда шкала непосредственного *количественного* измерения степени проявления этого свойства в объекте ему неизвестна (в том числе по причине объективного отсутствия таковой) или имеет условный смысл и интересует его только как *вспомогательное средство для последующего ранжирования рассматриваемых объектов*. К подобным ситуациям относится рассмотрение таких переменных, как «интегральный (сводный) показатель эффективности функционирования социально-экономической системы» (специалиста, предприятия, научно-производственного объединения и т. п.), «качество (мера оптимальности) структуры потребительского бюджета семьи», «качество жилищных условий семьи», «степень прогрессивности предлагаемого проекта решения социально-экономической, технической или другой проблемы» и т. п.

Таким образом, в отличие от статистического анализа k -го ($k = 0, 1, 2, \dots, p$) *количественного* признака $x^{(k)}$, когда в результате его измерения (наблюдения) на объектах мы могли каждому статистически обследованному объекту O_i поставить в соответствие некоторую *измеренную в физически интерпретируемой* шкале числовую характеристику $x_i^{(k)}$, результатом измерения *порядковой* переменной является приписывание каждому из обследованных объектов некоторой *условной числовой метки*, обозначающей место этого объекта в ряду из всех n анализируемых объектов, упорядоченном по убыванию степени проявления в них k -го изучаемого свойства. В этом случае $x_i^{(k)}$ называют *рангом* i -го объекта по k -му признаку.

Процесс упорядочения объектов $O_1, O_2, \dots, O_n$ производится либо с использованием *экспертной информации*, т. е. с привлечением экспертов, либо *формализованно* — путем перехода от исходного ряда наблюдений некоторого вспомогательного (косвенного, частного) количественного признака к соответствующему *вариационному ряду* (см. п. 6.3).

11.3.1. Исходные статистические данные (таблица или матрица рангов типа «объект-свойство»)

Итак, в результате измерения $p + 1$ порядковых переменных $x^{(0)} \equiv y, x^{(1)}, \dots, x^{(p)}$ на каждом из n анализируемых объектов $O_1, O_2, \dots, O_n$ мы получаем таблицу (матрицу) исходных данных следующего вида (табл. 11.1).

В этой таблице элемент $x_i^{(k)}$ задает порядковое место (*ранг*), которое занимает объект O_i в ряду всех статистически обследованных объектов, упорядоченном по убыванию степени проявления k -го анализируемого свойства (т. е. по переменной $x^{(k)}$).

Очевидно, если рассмотреть *столбец* с номером k этой таблицы ($k = 0, 1, \dots, p$), то он будет представлять перестановку из n элементов, а именно перестановку из n натуральных чисел $1, 2, \dots, n$, определяющую порядковые места объектов $O_1, O_2, \dots, O_n$ в ряду, упорядоченном по свойству $x^{(k)}$.

Замечание о случаях неразличимости рангов («объединенные ранги»). При упорядочении объектов по какому-либо свойству $x^{(k)}$ ($k = 0, 1, \dots, p$) могут встретиться ситуации, когда два объекта или целая группа их оказываются неразличимыми с точки зрения степени проявления в них этого свойства. Тогда каждому из объектов этой однородной группы приписывается ранг, равный среднему арифметическому значению тех мест, которые они делят, а полученные таким образом ранги принято называть «*объединенными*» (или «*связными*»). Так, например, упорядочивая семь альтернативных проектов A, B, C, D, E, F, G перспективного развития некоторой подотрасли с точки зрения их народнохозяйственной эффективности, эксперт поставил на 1-е место проект C , на 2-е — проект A , далее располагал проекты B, D и E , которые считал неразличимыми (равноценными) по эффективности, а последнее место отвел проектам F и G . Тогда соответствующий столбец таблицы «объект-свойство» будет состоять из следующих компонент:

$$x_A = 2; \quad x_B = \frac{3 + 4 + 5}{3} = 4; \quad x_C = 1; \quad x_D = x_E = \frac{3 + 4 + 5}{3} = 4;$$

$$x_F = x_G = \frac{6 + 7}{2} = 6,5.$$

Мы видим, что появление объединенных рангов может привести к *дробным значениям* рангов, составляющих массив исходных статистических данных (значения рангов, соответствующие 6-му и 7-му проектам). При отсутствии объединенных рангов область возможных значений пере-

менных $x^{(k)}$, очевидно, ограничивается множеством первых n чисел натурального ряда, где n — число сравниваемых объектов.

Таблица 11.1

Порядковый номер объекта («объект»)	Порядковый номер исследуемой переменной («свойство»)						
	0	1	3	...	k	...	p
1	$x_1^{(0)}$	$x_1^{(1)}$	$x_1^{(2)}$	...	$x_1^{(k)}$	...	$x_1^{(p)}$
2	$x_2^{(0)}$	$x_2^{(1)}$	$x_2^{(2)}$	...	$x_2^{(k)}$	...	$x_2^{(p)}$
...	...	...	...	...	...	...	...
i	$x_i^{(0)}$	$x_i^{(1)}$	$x_i^{(2)}$	...	$x_i^{(k)}$	...	$x_i^{(p)}$
...	...	...	...	...	...	...	...
n	$x_n^{(0)}$	$x_n^{(1)}$	$x_n^{(2)}$	...	$x_n^{(k)}$	...	$x_n^{(p)}$

Мы увидим далее, что наличие объединенных рангов несколько усложняет вычислительные процедуры, связанные со статистическим анализом соответствующих корреляционных характеристик.

11.3.2. Понятие ранговой корреляции

Под *ранговой корреляцией* понимается статистическая связь между порядковыми переменными. В статистической практике эта связь анализируется на основании исходных статистических данных, представленных упорядочениями (ранжировками) n рассматриваемых объектов по разным свойствам (см. столбцы табл. 11.1). Есть ли хоть какая-то согласованность (или связь) между упорядочением анализируемых объектов по свойству $x^{(k)}$ и упорядочением тех же объектов по другому свойству $x^{(j)}$? Можно ли измерить и проанализировать совокупную статистическую связь, существующую между ранжировками одних и тех же объектов $O_1, O_2, \dots, O_n$, полученными в соответствии со степенью проявления в них сначала свойства $x^{(k_1)}$ (1-й способ упорядочения), затем — свойства $x^{(k_2)}$ (2-й способ упорядочения)? Таким образом, речь идет о системе понятий и методов, позволяющих измерять и анализировать статистическую связь, существующую между двумя или несколькими ранжировками одного и того же конечного множества объектов $O_1, O_2, \dots, O_n$.

Система этих понятий и методов и составляет раздел математической статистики, который принято называть анализом *ранговых корреляций*.

ляций. Методы ранговой корреляции широко используются, в частности, при организации и статистической обработке различного рода систем экспертных обследований.

11.3.3. Основные задачи статистического анализа связей между ранжировками

Предположим, мы ввели измерители парной и множественной ранговой статистической связи (см. ниже, пп. 11.3.4, 11.3.5 и 11.3.8). Тогда, опираясь на эти характеристики, исследователь чаще всего пытается решить следующие три основные задачи статистического анализа структуры и характера связей, существующих между изучаемыми порядковыми переменными.

Задача А: анализ структуры имеющейся совокупности упорядочений $X^{(k)} = (x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)})^T$, $k = 0, 1, \dots, p$. Интерпретируя каждое упорядочение $X^{(k)}$ как точку в n -мерном пространстве, можно представить, например, три наиболее характерных типа такой структуры: 1) анализируемые точки *равномерно* разбросаны по всей области своих возможных значений (определяемой неравенствами $1 \leq x_i^{(k)} \leq n$, $i = 1, 2, \dots, n$), что означает отсутствие какой-либо связи или согласованности в представляемых ими ранжировках; 2) расположение $p + 1$ точек таково, что часть из них образует ядро из близко лежащих друг от друга точек («сгусток»), а остальные произвольно разбросаны относительно этого ядра. В этом случае существование ядра обеспечивает наличие подмножества согласованных переменных; 3) анализируемые точки — ранжировки располагаются в пространстве несколькими относительно далеко отстоящими друг от друга ядрами («сгустками»), что означает наличие нескольких подмножеств переменных таких, что переменные внутри одного подмножества обнаруживают высокую статистическую взаимосвязь, тогда как согласованности между переменными, взятыми из разных таких подсовокупностей, практически не существует.

Задача В: анализ интегральной (совокупной) согласованности рассматриваемых переменных и их условная ранжировка по критерию степени тесноты связи каждой из них с остальными переменными. Подобные задачи возникают, например, при исследовании степени согласованности мнений группы экспертов и при попытках условного упорядочения последних по их компетентности. В основе этого анализа лежит расчет коэффициента совокупной согласованности — *коэффициента конкордации* для различных комбинаций исследуемых переменных (см. п. 11.3.8).

Задача С: построение единого группового упорядочения объек-

тов на основе совокупности согласованных упорядочений «ядра» (или нескольких групповых упорядочений — при наличии нескольких «ядер»). Решение этой задачи сводится к построению такого упорядочения, которое было бы, в определенном смысле, наиболее близким к каждому из упорядочений заданной совокупности — «ядра». Именно с такой задачей сталкивается, например, исследователь, желающий установить *неизвестное истинное упорядочение* заданной совокупности объектов по имеющемуся в его распоряжении набору экспертных ранжировок тех же объектов. Для построения единого (группового) варианта упорядочения $X^{(ед)}$ часто используют в качестве ранга $x_i^{(ед)}$ объекта O_i среднее арифметическое или медиану имеющихся базовых рангов $x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)}$ этого объекта. Обоснование способа построения единого варианта упорядочения может быть получено, например, в рамках подхода, который опирается на меру близости между ранжировками (определяется ранжировка $X^{(ед)}$, наименее удаленная, в смысле введенной меры близости, от всех ранжировок $X^{(1)}, X^{(2)}, \dots, X^{(p)}$ базовой совокупности). Задача C может быть сформулирована и как задача наилучшего (в определенном смысле) восстановления ранжировки $X^{(0)}$, связанной с результирующей переменной $y \equiv x^{(0)}$, по ранжировкам $X^{(1)}, X^{(2)}, \dots, X^{(p)}$, индуцируемым соответственно объясняющими переменными $x^{(1)}, x^{(2)}, \dots, x^{(p)}$. В такой формулировке ее называют также *задачей регрессии на порядковые (ординальные) переменные*.

11.3.4. Ранговый коэффициент корреляции Спирмэна

Для измерения степени тесноты связи между ранжировками $X^{(k)} = (x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)})^T$ и $X^{(j)} = (x_1^{(j)}, x_2^{(j)}, \dots, x_n^{(j)})^T$ К. Спирмэн еще в 1904 г. предложил показатель

$$\hat{\tau}_{kj}^{(s)} = 1 - \frac{6}{n^3 - n} \sum_{i=1}^n (x_i^{(k)} - x_i^{(j)})^2, \quad (11.42)$$

названный впоследствии *ранговым коэффициентом корреляции Спирмэна*. Прямым подсчетом нетрудно убедиться, что для *совпадающих* ранжировок (т.е. при $x_i^{(k)} = x_i^{(j)}$ для всех $i = 1, 2, \dots, n$) $\hat{\tau}_{kj}^{(s)} = 1$ а для *противоположных* (т.е. при $x_i^{(k)} = n - x_i^{(j)} + 1$, $i = 1, 2, \dots, n$) — $\hat{\tau}_{kj}^{(s)} = -1$. Можно показать, что во всех остальных случаях $|\hat{\tau}_{kj}^{(s)}| < 1$.

Формула (11.42) пригодна лишь в случае отсутствия объединенных рангов в обеих исследуемых ранжировках. Для ее распространения на об-

ший случай определим для каждой (k -й) ранжировки $X^{(k)}$ ($k = 0, 1, \dots, p$) величину

$$T^{(k)} = \frac{1}{12} \sum_{t=1}^{m^{(k)}} [(n_t^{(k)})^3 - n_t^{(k)}], \quad (11.43)$$

где $m^{(k)}$ — число групп неразличимых рангов у переменной $x^{(k)}$, а $n_t^{(k)}$ — число элементов (рангов), входящих в t -ю группу неразличимых рангов (в частном случае отсутствия объединенных рангов имеем $m^{(k)} = n$, $n_1^{(k)} = n_2^{(k)} = \dots = n_n^{(k)} = 1$ и соответственно $T^{(k)} = 0$; кроме того, группы неразличимых рангов, состоящие из единственного элемента, по существу, не участвуют в расчете величины $T^{(k)}$).

Тогда ранговый коэффициент корреляции Спирмэна между ранжировками $X^{(k)}$ и $X^{(j)}$ следует вычислять по формуле

$$\tau_{kj}^{(s)} = \frac{\frac{1}{6}(n^3 - n) - \sum_{i=1}^n (x_i^{(k)} - x_i^{(j)})^2 - T^{(k)} - T^{(j)}}{\sqrt{\left[\frac{1}{6}(n^3 - n) - 2T^{(k)}\right] \left[\frac{1}{6}(n^3 - n) - 2T^{(j)}\right]}}. \quad (11.44)$$

Если $T^{(k)}$ и $T^{(j)}$ являются небольшими относительно $\frac{1}{6}(n^3 - n)$ величинами, то можно воспользоваться приближенным соотношением (а при $T^{(k)} = T^{(j)}$ оно точное)

$$\tau_{kj}^{(s)} = 1 - \frac{\sum_{i=1}^n (x_i^{(k)} - x_i^{(j)})^2}{\frac{1}{6}(n^3 - n) - (T^{(k)} + T^{(j)})}. \quad (11.44')$$

Правда, при этом же условии (относительная малость $T^{(k)} + T^{(j)}$ по сравнению с $\frac{1}{6}(n^3 - n)$) и приближенная формула (11.42) дает хорошую точность.

Пример 11.5. Два эксперта проранжировали 10 предложенных им проектов реорганизации научно-производственного объединения (НПО) с точки зрения их эффективности (при заданных ресурсных ограничениях). Пронумеровав проекты в порядке ранжировки 1-го эксперта, получаем в качестве исходных данных: $X^{(1)} = (1; 2; 3; 4; 5; 6; 7; 8; 9; 10)^T$; $X^{(2)} = (2; 3; 1; 4; 6; 5; 9; 7; 8; 10)^T$.

Вычисления по формуле (11.42) дают:

$$\tau_{12}^{(s)} = 1 - \frac{6}{1000 - 10} (1 + 1 + 2^2 + 0 + 1^2 + 1^2 + 2^2 + 1 + 1 + 0) = 1 - \frac{6}{990} \cdot 14 = 0,915,$$

что свидетельствует о существенной положительной ранговой связи между исследуемыми переменными.

Пример 11.6. Десять однородных предприятий подотрасли были проранжированы вначале по степени прогрессивности их оргструктур (признак $x^{(1)}$), а затем — по эффективности их функционирования в отчетном году (признак $x^{(2)}$). В результате были получены следующие две ранжировки: $X^{(1)} = (1; 2,5; 2,5; 4,5; 4,5; 6,5; 6,5; 8; 9,5; 9,5)^T$; $X^{(2)} = (1; 2; 4,5; 4,5; 4,5; 4,5; 8; 8; 8; 10)^T$.

В первой ранжировке имеем четыре группы неразличимых рангов, число элементов в которых больше единицы, а во второй ранжировке — две такие группы. В соответствии с формулой (11.43) получаем:

$$T^{(1)} = \frac{1}{12} [(2^3 - 2) + (2^3 - 2) + (2^3 - 2) + (2^3 - 2)] = \frac{24}{12} = 2,00;$$

$$T^{(2)} = \frac{1}{12} [(4^3 - 4) + (3^3 - 3)] = 7,42.$$

Точная формула (11.44) дает $\hat{r}_{12}^{(s)} = 0,917$. Вычисление этого же коэффициента корреляции по *приближенным* формулам (11.42) и (11.44') дает, соответственно, значения 0,921 и 0,917. Все эти результаты оказываются совпадающими при округлении до второго десятичного знака.

11.3.5. Ранговый коэффициент корреляции Кендалла

Другой широко используемой характеристикой тесноты статистической связи между двумя упорядочениями является ранговый коэффициент корреляции Кендалла, определяемый соотношением

$$\hat{r}_{kj}^{(K)} = 1 - \frac{4\nu(X^{(k)}, X^{(j)})}{n(n-1)}, \quad (11.45)$$

где $\nu(X^{(k)}, X^{(j)})$ — минимальное число обменов соседних элементов последовательности $X^{(j)}$, необходимое для приведения ее к упорядочению $X^{(k)}$. Очевидно, величина $\nu(X^{(k)}, X^{(j)})$ симметрична относительно своих аргументов, так что с равным правом можно говорить о минимальном числе «соседских обменов» элементов последовательности $X^{(k)}$, необходимом для приведения к виду $X^{(j)}$.

Из (11.45) сразу следует, что при совпадающих ранжировках $X^{(k)}$ и $X^{(j)}$ $\hat{r}_{kj}^{(K)} = 1$ (так как $\nu(X^{(k)}, X^{(j)}) = 0$), а при противоположных (т. е. при $x_i^{(k)} = n - x_i^{(j)} + 1$, $i = 1, 2, \dots, n$, так что $\nu(X^{(k)}, X^{(j)}) = \frac{1}{2}n(n-1)$) — $\hat{r}_{jk}^{(K)} = -1$. Нетрудно показать, что во всех остальных случаях $|\hat{r}_{kj}^{(K)}| < 1$.

Вычисление $\hat{\tau}_{kj}^{(K)}$ связано с необходимостью подсчета величины $\nu(X^{(k)}, X^{(j)})$ и, следовательно, является более *трудоемким*, чем вычисление $\tau_{kj}^{(S)}$. Однако, во-первых, коэффициент Кендалла обладает некоторыми преимуществами по сравнению с коэффициентом Спирмэна, главные из них: а) относительно большая продвинутость в исследовании его статистических свойств и, в частности, его выборочного распределения (см. ниже); б) возможность его использования и в частной («очищенной») корреляции рангов; в) большие удобства его пересчета при добавлении к n статистически обследованным объектам новых, т.е. при удлинении анализируемых ранжировок: для вычисления нового значения рангового коэффициента корреляции приходится переранжировать значительную часть объектов, что в случае $\tau_{kj}^{(S)}$ означает необходимость пересчета разностей $x_i^{(k)} - x_i^{(j)}$; при вычислении же $\tau_{ij}^{(K)}$ значения рангов не играют никакой роли, важно лишь число необходимых «соседских обменов», которое при добавлении новых объектов подсчитывается рекуррентным способом (к старому значению $\nu(X^{(k)}, X^{(j)})$ может быть лишь дополнен некоторый «добавок»).

Во-вторых, можно воспользоваться рекомендациями, упрощающими подсчет числа $\nu(X^{(k)}, X^{(j)})$ как при ручном, так и при машинном счете.

Так, при ручном счете полезным оказывается известный факт тождественного совпадения величин $\nu(X^{(k)}, X^{(j)})$ и $I(X^{(k)}, X^{(j)})$, где *число инверсий* $I(X^{(k)}, X^{(j)})$ — это просто число расположенных в неодинаковом порядке пар элементов последовательностей $X^{(k)}$ и $X^{(j)}$, являющееся естественной мерой нарушения порядка объектов в одной последовательности относительно другой. Для удобства подсчета $I(X^{(k)}, X^{(j)})$ перенумеруем объекты в порядке, определяемом рангами последовательности $X^{(k)}$. Тогда анализируемые ранжировки $X^{(k)}, X^{(j)}$ соответствующим образом видоизменяются, т.е. преобразуются к виду соответственно $\tilde{X}^{(k)}, \tilde{X}^{(j)}$, где $\tilde{X}^{(k)} = (1, 2, \dots, n)^T$; $\tilde{X}^{(j)} = (\tilde{x}_1^{(j)}, \tilde{x}_2^{(j)}, \dots, \tilde{x}_n^{(j)})^T$, а число инверсий $I(X^{(k)}, X^{(j)}) \equiv I(\tilde{X}^{(k)}, \tilde{X}^{(j)})$, а следовательно, и величина $\nu(X^{(k)}, X^{(j)})$ определяется по формуле

$$\nu(X^{(k)}, X^{(j)}) = I(X^{(k)}, X^{(j)}) = \sum_{q=1}^{n-1} \sum_{l=q+1}^n \nu_{ql}^{(j,k)}, \quad (11.46)$$

где

$$\nu_{ql}^{(j,k)} = \begin{cases} 1, & \text{если } \tilde{x}_q^{(j)} > \tilde{x}_l^{(j)} \quad (\text{т. е. нарушен порядок} \\ & \text{последовательности } \tilde{X}^{(k)}); \\ 0 & - \text{ в противоположном случае.} \end{cases}$$

Легко подсчитать, что число инверсий $I(X^{(k)}, X^{(j)})$ может меняться от 0 (что соответствует случаю совпадающих ранжировок) до $\frac{1}{2}n(n-1)$ (что соответствует случаю противоположных ранжировок).

Формулы (11.45)–(11.46) пригодны для подсчета $\hat{\tau}_{kj}^{(K)}$ лишь в случае отсутствия объединенных рангов в обеих исследуемых ранжировках. Соответствующее «подправленное» значение $\hat{\tau}_{kj}^{*(K)}$ при наличии объединенных рангов в анализируемых упорядочениях будет определяться соотношением

$$\hat{\tau}_{kj}^{*(K)} = \frac{\hat{\tau}_{kj}^{(K)} - \frac{2(U^{(k)} + U^{(j)})}{n(n-1)}}{\sqrt{\left(1 - \frac{2U^{(k)}}{n(n-1)}\right) \left(1 - \frac{2U^{(j)}}{n(n-1)}\right)}}, \quad (11.45')$$

в котором коэффициент $\hat{\tau}_{kj}^{(K)}$ вычисляется по формуле (11.45)–(11.46), а «поправочные» величины $U^{(l)}$ определяются соотношением

$$U^{(l)} = \frac{1}{2} \sum_{t=1}^{m^{(l)}} n_t^{(l)} (n_t^{(l)} - 1), \quad l = k, j \quad (11.47)$$

(смысл величин $m^{(l)}$ и $n_t^{(l)}$ определен выше, см. (11.43)).

Для пояснения работоспособности формул (11.45)–(11.47) вернемся к примерам 11.5 и 11.6.

Анализ степени согласованности ранжировок двумя экспертами десяти проектов реорганизации НПО (пример 11.5), осуществленный с использованием формул (11.45)–(11.46), дает:

$$\begin{aligned}
\nu_{12} &= 0; & \nu_{13} &= 1, \nu_{14} = \nu_{15} = \nu_{16} = \nu_{17} = \nu_{18} = \nu_{19} = \nu_{1.10} = 0; \\
\nu_{23} &= 1; & \nu_{24} &= \nu_{25} = \nu_{26} = \nu_{27} = \nu_{28} = \nu_{29} = \nu_{2.10} = 0; \\
\nu_{34} &= \nu_{35} = \nu_{36} = \nu_{37} = \nu_{38} = \nu_{39} = \nu_{3.10} = 0; \\
\nu_{45} &= \nu_{46} = \nu_{47} = \nu_{48} = \nu_{49} = \nu_{4.10} = 0; \\
\nu_{56} &= 1; & \nu_{57} &= \nu_{58} = \nu_{59} = \nu_{5.10} = 0; \\
\nu_{67} &= \nu_{68} = \nu_{69} = \nu_{6.10} = 0; \\
\nu_{78} &= 1; & \nu_{79} &= 1; & \nu_{7.10} &= 0; \\
\nu_{89} &= \nu_{8.10} = 0; \\
\nu_{9.10} &= 0.
\end{aligned}$$

Таким образом, $\nu(X^{(1)}, X^{(2)}) = 1 + 1 + 0 + 0 + 1 + 0 + 2 + 0 + 0 = 5$.

Соответственно

$$\hat{\tau}_{12}^{(K)} = 1 - \frac{4 \cdot 5}{10 \cdot 9} = 1 - 0,222 = 0,778$$

(напомним, что коэффициент Спирмэна в этом примере был равным 0,915).

При вычислении рангового коэффициента корреляции Кендалла в примере 11.6 следует воспользоваться формулой (11.45'), так как исследуемые ранжировки содержат объединенные ранги. Используя результаты расчета величин $m^{(1)} = 4$, $m^{(2)} = 2$, $n_1^{(1)} = n_2^{(1)} = n_3^{(1)} = n_4^{(1)} = 2$, $n_1^{(2)} = 4$, $n_2^{(2)} = 3$, получаем (в соответствии с (2.8)):

$$U^{(1)} = \frac{1}{2}(2 + 2 + 2 + 2) = 4; \quad U^{(2)} = \frac{1}{2}(4 \cdot 3 + 3 \cdot 2) = 9.$$

Обращаясь теперь к формуле (2.6'), имеем:

$$\hat{\tau}_{12}^{*(K)} = \frac{1 - \frac{26}{90}}{\sqrt{(1 - \frac{8}{90})(1 - \frac{18}{90})}} = 0,833$$

(напомним, что соответствующий коэффициент Спирмэна был равен 0,917).

11.3.6. Обобщенная формула для парного коэффициента корреляции и связь между коэффициентами Спирмэна и Кендалла

Для удобства стандартной вычислительной реализации системы алгоритмов корреляционного анализа полезно ввести некоторый *обобщенный* прием вычисления парных корреляционных характеристик, определенный для любой двумерной системы n наблюдений

$$\begin{pmatrix} X^{(k)'} \\ X^{(j)'} \end{pmatrix} = \begin{pmatrix} x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)} \\ x_1^{(j)}, x_2^{(j)}, \dots, x_n^{(j)} \end{pmatrix}. \quad (11.48)$$

С этой целью определим некоторое правило, в соответствии с которым каждой паре $(x_{i_1}^{(l)}, x_{i_2}^{(l)})$ компонент вектора $X^{(l)}$ ($l = k, j$) ставится в соответствие число («метка») $a_{i_1 i_2}^{(l)}$, причем это правило должно обладать свойством *отрицательной симметричности* (т.е. $a_{i_1 i_2}^{(l)} = -a_{i_2 i_1}^{(l)}$) и *центрированности* (т.е. $a_{ii}^{(l)} = 0$ при всех $l = k, j$ и всех $i = 1, 2, \dots, n$). Тогда *обобщенный коэффициент корреляции* $r^{(об)}$ переменных $x^{(k)}$ и $x^{(j)}$ определяется формулой

$$\hat{r}_{kj}^{(об)} = \frac{\sum_{i_1=1}^n \sum_{i_2=1}^n a_{i_1 i_2}^{(k)} a_{i_1 i_2}^{(j)}}{\sqrt{\sum_{i_1=1}^n \sum_{i_2=1}^n (a_{i_1 i_2}^{(k)})^2 \cdot \sum_{i_1=1}^n \sum_{i_2=1}^n (a_{i_1 i_2}^{(j)})^2}}. \quad (11.49)$$

Легко видеть, что практически все введенные нами характеристики парной корреляционной связи могут быть получены как частные случаи формулы (11.49) при соответствующем выборе правила приписывания числовых «меток» $a_{i_1 i_2}$. Действительно:

а) положив $a_{i_1 i_2}^{(l)} = x_{i_1}^{(l)} - x_{i_2}^{(l)}$, $l = k, j$, получаем формулу для *обычного парного коэффициента корреляции* $\hat{r}_{kj}$, если $x_i^{(l)}$ — значение l -количественной переменной в i -м наблюдении (см. п. 11.2.2, формулу (11.13')), и формулу для *рангового коэффициента корреляции* Спирмэна $\hat{r}_{kj}^{(S)}$, если $x_i^{(l)}$ — ранг i -го объекта в ряду, упорядоченном по порядковой переменной $x^{(l)}$ (см. формулу (11.42));

б) положив

$$a_{i_1 i_2}^{(l)} = \begin{cases} +1, & \text{если } x_{i_1}^{(l)} < x_{i_2}^{(l)}; \\ 0, & \text{если } x_{i_1}^{(l)} = x_{i_2}^{(l)}; \\ -1, & \text{если } x_{i_1}^{(l)} > x_{i_2}^{(l)}, \end{cases}$$

получаем формулы (11.45) и (11.45') для рангового коэффициента корреляции Кендалла $\hat{\tau}_{kj}^{(K)}$, если под $x_i^{(l)}$ понимать ранг i -го объекта в l -м упорядочении.

Заметим, что значения ранговых корреляционных характеристик $\hat{\tau}_{kj}^{(K)}$ и $\hat{\tau}_{kj}^{(S)}$ довольно тесно связаны одно с другим. Это следовало ожидать, так как *обе характеристики являются линейными функциями от числа инверсий*, имеющих в сравнении последовательностей $X^{(k)}$ и $X^{(j)}$: различие этих функций состоит в том, что при подсчете коэффициента Спирмэна инверсиям более отдаленных (по величине) друг от друга элементов приписываются большие веса. Между масштабами шкал, в которых измеряют корреляцию коэффициенты $\hat{\tau}^{(S)}$ и $\hat{\tau}^{(K)}$, нет простого соотношения. Однако уже при умеренно больших значениях n ($n \geq 10$) и при условии, что абсолютные величины значений этих коэффициентов не слишком близки к единице, их связывает следующее простое приближенное соотношение $\hat{\tau}^{(S)} \approx 1,5\hat{\tau}^{(K)}$.

11.3.7. Статистические свойства выборочных характеристик парной ранговой связи

До сих пор речь шла о *выборочных* характеристиках ранговой связи. Попробуем ответить на вопрос: как точно эти выборочные характеристики (определенные, в частности, формулами (11.42)–(11.47)) оценивают соответствующие истинные (*теоретические*) значения?

Для этого в первую очередь следует пояснить, что в данном случае понимается под теоретическими характеристиками.

Представим себе сначала *конечную* генеральную совокупность, состоящую из N объектов $O_1, O_2, \dots, O_N$, каждый из которых снабжен двумя порядковыми номерами: $O_i \leftrightarrow (x_i^{(k)}, x_i^{(j)})$, $i = 1, 2, \dots, N$, где $x_i^{(l)}$ означает место объекта O_i в общем ряду всех N объектов, упорядоченном по степени выраженности свойства $x^{(l)}$ ($l = k, j$). Будем полагать, что статистически обследованное множество объектов $O_{i_1}, O_{i_2}, \dots, O_{i_n}$ образуется как случайная выборка объема n , взятая из совокупности $O_1, O_2, \dots, O_N$ ($n \ll N$).

Определим *теоретические* (истинные) значения коэффициентов $\tau_{kj}^{(S)}$ и $\tau_{kj}^{(K)}$ соответственно теми же соотношениями (11.42) (или (11.44)) и (11.45) (или (11.45')), что и выборочные с заменой объема выборки n объемом генеральной совокупности N . При работе с выборкой производится *естественная перенумерация* объектов и их рангов, не меняющая их упорядоченности в генеральной совокупности ни по одной из переменных.

В дальнейшем нас будет интересовать, как сильно могут отличаться выборочные значения $\hat{\tau}^{(S)}$ и $\hat{\tau}^{(K)}$ от соответствующих теоретических, в том числе в так называемых *асимптотических ситуациях*, т.е. при $N \rightarrow \infty$ и $n(N) \rightarrow \infty$.

Проверка статистически значимого отличия от нуля ранговых корреляционных характеристик осуществляется при «не слишком малых» n ($n > 10$) и заданном уровне значимости критерия α с помощью неравенства

$$|\hat{\tau}^{(S)}| > t_{\frac{\alpha}{2}}(n-2) \sqrt{\frac{1 - (\hat{\tau}^{(S)})^2}{n-2}}; \quad (11.50)$$

$$|\hat{\tau}^{(K)}| > u_{1-\alpha/2}(n-2) \sqrt{\frac{2(2n+5)}{9n(n-1)}}, \quad (11.51)$$

в которых $t_q(\nu)$ и u_q , как и прежде, соответственно 100 q %-ная точка $t(\nu)$ -и q -квантиль стандартного нормального распределения (см. табл. в Приложении 1). Выполнение неравенств (11.50) и (11.51) сигнализирует о необходимости отвергнуть гипотезу об отсутствии статистически значимой ранговой корреляционной связи. В случае небольших объемов выборок ($4 \leq n \leq 10$) статистическая проверка гипотезы об отсутствии ранговой корреляционной связи производится с помощью специальных таблиц (см. Приложение 1).

Таблица значений вспомогательной величины S_C позволяет при малых n ($n = 4, 5, \dots, 10$) построить то пороговое значение $\tau_{\max}^{(S)}$, при превышении которого (по абсолютной величине) коэффициентом Спирмэна $\hat{\tau}^{(S)}$ следует признать наличие статистически значимой связи между анализируемыми переменными. Задавшись уровнем значимости критерия α и числом сравниваемых объектов n , определяем из таблицы величину $S_C = S_C(n, Q)$, соответствующую нашему n и значению $Q = \alpha/2$ (или приблизительно равному $\alpha/2$). Тогда

$$\tau_{\max}^{(S)} = \frac{2S_C(n, Q)}{K_n} - 1, \quad (11.52)$$

где $K_n = \frac{1}{3}(n^3 - n)$ (значения этой вспомогательной константы приведены в последней строке таблицы).

Так, в примере 11.5 для уровня значимости $\alpha = 0,06$ имеем: $n = 10$; $Q = 0,03$; $S_C = S_C(10; 0,3) = 268$; $K_{10} = 330$, так что в соответствии с (11.52)

$$\tau_{\max}^{(S)} = \frac{2 \cdot 268}{330} - 1 = 0,624.$$

Поскольку выборочное значение рангового коэффициента корреляции Спирмэна $\tau^{(S)}$ в этом примере значительно превосходит пороговое значение ($\hat{\tau}^{(S)} = 0,915 > 0,624$), то гипотеза об отсутствии корреляционной связи отвергается.

И наконец, в таблице значений вспомогательной величины S_K приведены значения величин S_K , позволяющих вычислить (при малых $n = 4, 5, \dots, 10$) то пороговое значение $\tau_{\max}^{(K)}$, при превышении которого (по абсолютной величине) коэффициентом Кендалла следует признать наличие статистически значимой связи между анализируемыми переменными. Для этого поступают следующим образом: задавшись объемом выборки n и уровнем значимости критерия α , находят в столбце, соответствующем данному n , величину, равную (или приблизительно равную) $\alpha/2$; затем находят значение $S_K = S_K(n, \alpha)$ в левом столбце *той же самой строки* и вычисляют $\tau_{\max}^{(K)}$ по формуле

$$\tau_{\max}^{(K)} = \frac{2S_K(n, \alpha)}{n(n-1)}. \quad (11.53)$$

Если окажется, что $\hat{\tau}^{(K)} > \tau_{\max}^{(K)}$, то гипотеза об отсутствии ранговой корреляционной связи отвергается (связь статистически значима).

Так, в примере 11.5 при уровне значимости $\alpha = 0,06$ имеем: $n = 10$; $0,23 < \frac{\alpha}{2} < 0,36$; следовательно, $S_K = 22$ (оно лежит между 21 и 23), так что

$$\tau_{\max}^{(K)} = \frac{2 \cdot 22}{10 \cdot 9} = \frac{44}{90} = 0,489.$$

Поскольку $\tau^{(K)} = 0,733 > 0,489$, делается вывод о наличии статистически значимой корреляционной связи между исследуемыми переменными в данном примере.

Построение доверительных интервалов для неизвестных истинных значений ранговых коэффициентов корреляции возможно лишь приближенно и только при измерении ранговой корреляции с помощью коэффициента Кендалла. При этом используют (при $n > 10$ и значениях $\tau^{(K)}$, не слишком близких по абсолютной величине к единице) приближенный факт нормальности распределения величины $\hat{\tau}^{(K)}$ со средним значением $E\hat{\tau}^{(K)} \approx \tau^{(K)}$ и с дисперсией $D\hat{\tau}^{(K)}$, не превышающей величины $\frac{2}{n}[1 - (\tau^{(K)})^2]$. Можно утверждать, что с доверительной вероятностью, не меньшей заданного уровня P , истинное значение коэффициента Кендалла

$\tau^{(K)}$ заключено в пределах

$$\begin{aligned} \hat{\tau}^{(K)} - u_{\frac{1+p}{2}} \sqrt{\frac{2}{n} [1 - (\hat{\tau}^{(K)})^2]} < \tau^{(K)} < \hat{\tau}^{(K)} \\ + u_{\frac{1+p}{2}} \sqrt{\frac{2}{n} [1 - (\hat{\tau}^{(K)})^2]}, \end{aligned} \quad (11.54)$$

где u_q — q -квантиль стандартного нормального распределения.

11.3.8. Коэффициент конкордации (согласованности) как измеритель статистической связи между несколькими порядковыми переменными

До сих пор мы рассматривали корреляцию между двумя порядковыми переменными. Однако при решении основных задач $A - C$ статистического анализа ранговых связей (см. п. 11.3.3) возникает необходимость уметь измерить статистическую связь между *несколькими* (более чем двумя) переменными. С этой целью Кендаллом был предложен показатель $\widehat{W}(m)$, названный *коэффициентом конкордации* (или *согласованности*), вычисляемый по формуле¹

$$\widehat{W}(m) = \frac{12}{m^2(n^3 - n)} \sum_{i=1}^n \left(\sum_{j=1}^m x_i^{(k_j)} - \frac{m(n+1)}{2} \right)^2, \quad (11.55)$$

где m — число анализируемых порядковых переменных (сравниваемых упорядочений); n — число статистически обследованных объектов или длина ранжировки (объем выборки); $k_1, k_2, \dots, k_m$ — номера отобранных для анализа порядковых переменных (из исходной совокупности $x^{(0)}, x^{(1)}, x^{(2)}, \dots, x^{(p)}$), так что, очевидно, $m \leq p + 1$).

Нетрудно устанавливаются следующие свойства коэффициента конкордации:

а) $0 \leq \widehat{W} \leq 1$;

б) $\widehat{W} = 1$ тогда и только тогда, когда все m анализируемых упорядочений совпадают;

¹ Мы приводим здесь формулу для подсчета *выборочного* значения $\widehat{W}$ коэффициента конкордации W . Интерпретация и вычисление *теоретического* значения W непосредственно следуют из рассуждений, приведенных в п. 11.3.7 в связи с анализом статистических свойств выборочных парных ранговых коэффициентов корреляции.

в) если $m \geq 3$ и анализируемые ранжировки генерируются подобно случайному независимому m -кратному извлечению из множества всех n возможных упорядочений n объектов, то связи между ними нет и $W = 0$;

г) пусть $\bar{\tau}^{(S)}(m)$ — среднее значение коэффициента Спирмэна, подсчитанное по значениям $m(m-1)/2$ коэффициентов $\hat{\tau}_{ki k_j}^{(S)}$ ($i, j = 1, 2, \dots, m; i \neq j$), характеризующих ранговую связь между всеми возможными парами переменных $(x^{(k_i)}, x^{(k_j)})$ из анализируемого набора $(x^{(k_1)}, x^{(k_2)}, \dots, x^{(k_m)})$; тогда

$$\bar{\tau}^{(S)}(m) = \frac{m\widehat{W}(m) - 1}{m - 1}; \quad (11.56)$$

в частности, из (11.56) следует для случая $m = 2$, что

$$\widehat{W}(2) = \frac{1}{2}(\hat{\tau}_{k_1 k_2}^{(S)} + 1), \quad (11.56')$$

т.е. коэффициент конкордации, исчисленный для *двух* переменных, пропорционален введенному ранее парному ранговому коэффициенту корреляции Спирмэна.

То, что шкала измерения $W(m)$ не включает в себя отрицательных значений, объясняется следующим обстоятельством. В отличие от случая парных связей при анализе $m(m \geq 3)$ порядковых переменных противоположные понятия согласованности и несогласованности утрачивают прежнюю симметричность (относительно нуля); упорядочения, произведенные в соответствии с переменными $x^{(k_1)}, x^{(k_2)}, \dots, x^{(k_m)}$, могут полностью совпадать, но *не могут полностью не совпадать* в том смысле, который мы вкладывали в это понятие при $m = 2$.

Формула (11.55) получена (и справедлива) в предположении отсутствия объединенных рангов в каждом из анализируемых упорядочений. Если же таковые имеются, то формула должна быть модифицирована:

$$\widehat{W}(m) = \frac{\sum_{i=1}^n \left(\sum_{j=1}^m x_i^{(k_j)} - \frac{m(n+1)}{2} \right)^2}{\frac{1}{12} m^2 (n^3 - n) - m \sum_{j=1}^m T^{(k_j)}}, \quad (11.55')$$

где поправочный коэффициент $T^{(k_j)}$ (соответствующий переменной $x^{(k_j)}$) подсчитывается по формуле (11.43).

11.3.9. Проверка статистической значимости выборочного значения коэффициента конкордации

Как ведут себя выборочные значения $\widehat{W}(m)$ коэффициента конкордации при повторении выборок заданного объема n (из одной и той же генеральной совокупности) при отсутствии какой-либо связи между анализируемыми m переменными? Другими словами, нас интересует ответ на следующий вопрос. Предположим, что каждому объекту конечной *генеральной совокупности* (состоящей из N элементов) приписан какой-то определенный ранг по каждой из m рассматриваемых переменных. Так, например, если $m = 3$ и объекту O_i приписана тройка $(x_i^{(1)} = N; x_i^{(2)} = 1; x_i^{(3)} = 2)$, то это означает, что по переменной $x^{(1)}$ он стоит на последнем (N -м) месте в упорядоченном ряду *всех* объектов генеральной совокупности, по переменной $x^{(2)}$ — на первом и по переменной $x^{(3)}$ — на втором. Тогда по исходным данным $\{(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(m)})\}_{i=1, \dots, N}$ с помощью формулы (11.55) может быть вычислен теоретический (генеральный) коэффициент конкордации $W(m)$, характеризующий степень тесноты ранговой связи между переменными $x^{(1)}, x^{(2)}, \dots, x^{(m)}$. Однако исследователю известны значения $(x^{(1)}, x^{(2)}, \dots, x^{(m)})$ *лишь для части объектов* генеральной совокупности, а именно для случайной выборки объектов объема n ($n < N$). После естественной перенумерации рангов, сохраняющей правило упорядочения объектов, но переводящей масштаб измерения рангов в шкалу $(1, 2, \dots, n)$ (для этого минимальный из оказавшихся в выборке рангов по каждой переменной объявляется рангом, равным 1, следующий по величине — рангом, равным 2, и т. д.), может быть вычислен (по той же формуле (11.55)) *выборочный* коэффициент конкордации $\widehat{W}(m)$. Извлекая другую выборку объема n из той же самой генеральной совокупности, мы получим, вообще говоря, другое значение выборочного коэффициента $\widehat{W}(m)$ и т. д.

Спрашивается, как сильно могут отклоняться от нуля выборочные значения коэффициента конкордации $\widehat{W}(m)$ в ситуации, когда значение теоретического коэффициента конкордации $W(m)$ свидетельствует о полном отсутствии ранговой связи между анализируемыми переменными $x^{(1)}, x^{(2)}, \dots, x^{(m)}$? Для малых значений m и n ($2 \leq m \leq 20$, $3 \leq n \leq 7$) ответ на этот вопрос может быть получен с помощью таблицы значений величины S . Обозначенная в ней величина S есть не что иное, как

$$S = \sum_{i=1}^n \left(\sum_{j=1}^m x_i^{(k_j)} - \frac{m(n+1)}{2} \right)^2. \quad (11.57)$$

«Входами» в эту таблицу является тройка чисел (m, n, S) , «выходом» — вероятность того, что величина S может быть такой, какой она является в нашей выборке, или большей в условиях отсутствия связи переменных в генеральной совокупности. Если окажется, что эта вероятность меньше принятой нами величины уровня значимости критерия α (например, $\alpha = 0,05$), то гипотезу об отсутствии связи следует отвергнуть, т. е. признать статистическую значимость анализируемой связи. Таблица критических значений $W(m)$ построена несколько иначе. В ней при уровне значимости $\alpha = 0,05$ и в соответствии с «входами» (m, n) даны «критические» значения величины S , т. е. такие значения, при превышении которых следует отвергать гипотезу об отсутствии связей (признавать их статистическую значимость).

При $n > 7$ для проверки статистической значимости анализируемой связи следует воспользоваться фактом приближенной $\chi^2(n-1)$ -распределенности величины $m(n-1) \times \widehat{W}(m)$, справедливым в условиях отсутствия связи в генеральной совокупности (значение $\widehat{W}(m)$, как и прежде, подсчитывается по формуле (2.16) или (2.16')). Поэтому, если окажется, что

$$m(n-1)\widehat{W}(m) > \chi_\alpha^2(n-1), \quad (11.58)$$

то гипотеза об отсутствии ранговой связи между переменными $x^{(k_1)}, x^{(k_2)}, \dots, x^{(k_m)}$ должна быть отвергнута (с уровнем значимости критерия, равным α); в (11.58) величина $\chi_\alpha^2(n-1)$ — это 100 α %-ная точка χ^2 -распределения с $(n-1)$ -й степенью свободы.

Можно использовать и другой способ проверки статистической значимости исследуемой ранговой связи между несколькими переменными, основанный на том, что в условиях отсутствия таковой в генеральной совокупности распределение случайной величины $\frac{1}{2} \ln \frac{(m-1)\widehat{W}(m)}{1-\widehat{W}(m)}$ приближенно описывается Z -распределением Фишера с числом степеней свободы числителя $\nu_1 = n - 1 - \frac{2}{m}$ и знаменателя $\nu_2 = (m-1)\nu_1$.

Строгих рекомендаций по построению доверительных интервалов для истинного значения W в условиях *наличия* ранговых связей в исследуемой генеральной совокупности к настоящему времени не имеется.

Рассмотрим примеры, в которых реализуются приведенные выше рекомендации по статистическому анализу множественных ранговых связей.

Пример 11.7. Рассмотрим три порядковые переменные $(x^{(1)}, x^{(2)}, x^{(3)})$ и соответствующие им упорядочения десяти объектов:

$X^{(1)\top}$	1	4,5	2	4,5	3	7,5	6	9	7,5	10
$X^{(2)\top}$	2,5	1	2,5	4,5	4,5	8	9	6,5	10	6,5
$X^{(3)\top}$	2	1	4,5	4,5	4,5	4,5	8	8	8	10
Сумма	5,5	6,5	9	13,5	12	20	23	23,5	23,5	26,5

В соответствии с формулами (11.57), (11.43) имеем:

$$\begin{aligned}
 S &= \sum_{i=1}^{10} \left(\sum_{j=1}^3 x_i^{(j)} - \frac{3 \cdot 11}{2} \right)^2 = (-11)^2 + (-10)^2 + (-7,5)^2 \\
 &\quad + (-3)^2 + (-4,5)^2 + (3,5)^2 + (6,5)^2 + 7^2 + 9^2 + 10^2 = 591; \\
 T^{(1)} &= \frac{1}{12}(2^3 - 2)2 = 1; \\
 T^{(2)} &= \frac{1}{12}(2^3 - 2)2 = 1,5; \\
 T^{(3)} &= \frac{1}{12}(4^3 - 4 + 3^3 - 3) = 7.
 \end{aligned}$$

Следовательно, в соответствии с (11.55')

$$\widehat{W}(3) = \frac{591}{\frac{1}{12}3^2(10^2 - 10) - 3(1 + 1,5 + 7)} = \frac{591}{742,5 - 28,5} = 0,828.$$

Пример 11.8. Требуется проверить статистическую значимость множественной ранговой связи 28 переменных ($m = 28$), характеризуемой величиной выборочного коэффициента конкордации $\widehat{W}(28) = 0,08$, подсчитанного по 13 объектам ($n = 13$).

Воспользуемся фактом $\chi^2(12)$ -распределенности случайной величины $m(n-1)\widehat{W}(m)$, который имеет место (приближенно) в случае, если в исследуемой генеральной совокупности множественная ранговая связь отсутствует. Тогда критерий сводится к проверке неравенства (11.57). Задав-шись уровнем значимости критерия $\alpha = 0,05$, находим из таблиц значение 5%-ной точки χ^2 -распределения с 12 степенями свободы $\chi_{0,05}^2(12) = 21,026$. В то же время $m(n-1)\widehat{W}(m) = 28 \cdot 12 \cdot 0,08 = 27$.

Поскольку $m(n-1)\widehat{W}(m) > \chi_{0,05}^2(12)$, то оказалось, что *даже такого маленького числа, как 0,08, «хватило»* для того, чтобы объявить связь между 28 исследуемыми переменными статистически значимой.

11.4. Корреляционный анализ категоризованных переменных: таблицы сопряженности

Признак называют *категоризованным*, если его возможные «значения» описываются конечным числом состояний или градаций. Слово «значения» взято в кавычки, так как речь идет, как правило, не о числовых значениях, а лишь об определенных условных метках возможных состояний. Так, категоризованный признак «пол индивидуума» имеет две градации, категоризованный признак «социальное положение респондента» — столько градаций, сколько установлено социальных страт (слов) в данном обществе, категоризованный признак «уровень жилищных условий семьи» — столько градаций, сколько их определено в данном обследовании (например, при возможных ответах: «низкий», «удовлетворительный», «хороший» и «очень хороший» мы будем иметь четыре градации). Как мы видим, в класс категоризованных признаков попадают те номинальные¹ и порядковые (ординальные) переменные, возможные значения которых описаны заданным (*известным*) набором градаций.

11.4.1. Исходные статистические данные (таблицы сопряженности)

Мы ограничимся здесь задачей измерения *парных* статистических связей между категоризованными переменными. Формально исходные данные для пары категоризованных переменных будут иметь уже знакомый нам вид (9.1') таблицы «объект-свойство», в которой, правда, в качестве элементов $x_i^{(j)}$ ($j = 1, 2; i = 1, 2, \dots, n$) будут обозначены *условные метки (градации)* состояния объекта i по переменной j . Однако мы не можем работать с этими данными аналогично случаю количественных переменных: скажем, среднее значение или дисперсия «пола» или «социальной принадлежности» респондента лишены всякого смысла. Поэтому при статистическом анализе двух категоризованных переменных исходные данные преобразуют к виду таблицы перекрестных частот, называемой *двухвходовой таблицей сопряженности* признаков $x^{(1)}$ и $x^{(2)}$ (см. табл. 11.2) или просто таблицей сопряженности.

¹ Напомним, что случайная величина называется номинальной, если знание ее значений на статистически обследованных объектах позволяет разбить это множество на не поддающиеся упорядочению однородные по анализируемому свойству классы (см. п. 2.3)

Таблица 11.2

Градация признака $x^{(1)}$	Градация признака $x^{(2)}$						$n_{i.}$
	1	2	...	j	...	m_2	
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1m_2}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2m_2}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{im_2}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
m_1	$n_{m_1 1}$	$n_{m_1 2}$	...	$n_{m_1 j}$	...	$n_{m_1 m_2}$	$n_{m_1 .}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.m_2}$	n

В табл. 11.2 представлены результаты статистического обследования n объектов по признакам $x^{(1)}$ и $x^{(2)}$. В ней n_{ij} означает число объектов (из общего числа n обследованных), у которых «значение» признака $x^{(1)}$ оказалось зафиксированным на уровне i -й градации, а «значение» признака $x^{(2)}$ — на уровне j -й градации; $n_{i.} = \sum_{j=1}^{m_2} n_{ij}$ — число объектов, значение признака $x^{(1)}$ у которых оказалось зарегистрированным на уровне i -й градации (при отсутствии каких бы то ни было условий на «значения» признака $x^{(2)}$, а $n_{.j} = \sum_{i=1}^{m_1} n_{ij}$ — число объектов, значение признака $x^{(2)}$ у которых оказалось зарегистрированным на уровне j -й градации ($i = 1, 2, \dots, m_1$; $j = 1, 2, \dots, m_2$).

Переход к задаче измерения степени тесноты связи *несколько* (более двух) категоризованных переменных связан с необходимостью введения *многовходовых таблиц сопряженности* и более громоздких обозначений (количество нижних индексов у анализируемых часто приходится увеличивать до числа анализируемых переменных).

11.4.2. Основные

измерители степени тесноты статистической связи между двумя категоризованными переменными

В статистической теории и практике в рамках данной проблемы существует *целый спектр характеристик с.т.с.с.* Разные меры связи акцентируют внимание на разных аспектах взаимоотношений между переменными, давая в целом многоаспектную информацию о природе изучаемой

зависимости. Поэтому вряд ли целесообразно давать универсальные рекомендации по поводу выбора наиболее предпочтительной характеристики «на все случаи жизни». Выбор описанных ниже измерителей с.т.с.с. был основан на сочетании двух критериев: 1) *наибольшей распространенности* в практике статистических исследований; 2) *умении описать необходимые статистические свойства* этих измерителей.

Построение подавляющего большинства используемых в данной проблеме характеристик с.т.с.с. основано на одной общей идее: *характеристика должна принимать тем большее числовое значение, чем больше анализируемая ситуация отклоняется от гипотезы взаимной статистической независимости исследуемых переменных $x^{(1)}, x^{(2)}$* . Как мы знаем (см. п. 2.5.3), формализуется понятие статистической независимости условием представления двумерного закона распределения векторной случайной величины $X = (x^{(1)}, x^{(2)})^T$ в виде произведения частных законов распределения компонент этого вектора, т. е.

$$P\{x^{(1)} = x_i^{(1)0}; x^{(2)} = x_j^{(2)0}\} = P\{x^{(1)} = x_i^{(1)0}\}P\{x^{(2)} = x_j^{(2)0}\}, \quad (11.59)$$

где $x_i^{(1)0}$ и $x_j^{(2)0}$ — обозначения принятых в данном исследовании меток соответственно для i -й градации признака $x^{(1)}$ и j -й градации признака $x^{(2)}$.

Выборочный (эмпирический) эквивалент этого условия предусматривает замену участвующих в нем вероятностей соответствующими относительными частотами (с заменой, конечно, *точного* знака равенства на *приближенный*), т. е.

$$\frac{n_{ij}}{n} \approx \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n} \quad (11.59')$$

или

$$D_{ij} = \frac{n_{ij}}{n} - \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n} \approx 0. \quad (11.59'')$$

Поэтому чем больше от нуля будет отличаться просуммированная по i и j левая часть соотношения (11.59''), вычисленная по данным таблицы сопряженности 11.2, тем более высокую с.т.с.с. между анализируемыми переменными должны обозначать используемые измерители.

Характеристика X^2 квадратичной сопряженности признаков $x^{(1)}$ и $x^{(2)}$ определяется соотношением

$$X^2 = n \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{D_{ij}^2}{n_{i.} n_{.j}} = n \left(\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1 \right). \quad (11.60)$$

Его значение может меняться от нуля (при строгой статистической независимости переменных $x^{(1)}$ и $x^{(2)}$) до $+\infty$. Являясь, как и всякая функция от выборочных данных, величиной случайной, X^2 ведет себя (в предположении статистической независимости $x^{(1)}$ и $x^{(2)}$) приблизительно, — асимптотически по $n \rightarrow \infty$, — как χ^2 -распределенная случайная величина с числом степеней свободы, равным $(m_1 - 1)(m_2 - 1)$. Это дает нам право в случае

$$X^2 \geq \chi^2_{\alpha}((m_1 - 1)(m_2 - 1)) \quad (11.61)$$

заявить о том, что связь между анализируемыми переменными $x^{(1)}$ и $x^{(2)}$ является *статистически значимой* и она тем теснее, чем больше значение X^2 . Неудобство использования характеристики X^2 связано с тем, что ее верхняя граница стремится к бесконечности при возрастании объема выборки n . Поэтому вместо X^2 часто используют характеристику

$$C = \left[\frac{X^2}{n \min(m_1 - 1, m_2 - 1)} \right]^{1/2}, \quad (11.62)$$

которая называется **коэффициентом Крамера** и которая обладает более привычным для измерителей с.т.с.с. диапазоном изменения, а именно $0 \leq C \leq 1$. При этом нулевое значение C свидетельствует о строгой статистической независимости анализируемых признаков, а значение C , равное единице, — о возможности однозначного восстановления значения одной переменной по известному значению другой.

Информационная характеристика с.т.с.с. Y^2 признаков $x^{(1)}$ и $x^{(2)}$ также основана на мере отклонения от выполнения соотношения независимости (11.59'), только вместо разности левых и правых частей этого соотношения в ней используется их отношение. А именно измеритель Y^2 определяется соотношением

$$Y^2 = 2 \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} n_{ij} \ln \left(\frac{n_{ij}}{n_{i.} n_{.j} / n} \right). \quad (11.63)$$

Правую часть (11.63) можно преобразовать к виду, более удобному для вычислений

$$Y^2 = 2 \left(\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} n_{ij} \ln n_{ij} - \sum_{i=1}^{m_1} n_{i.} \ln n_{i.} - \sum_{j=1}^{m_2} n_{.j} \ln n_{.j} + n \ln n \right). \quad (11.63')$$

Эта характеристика обладает теми же свойствами, что и X^2 : ее значения варьируют от 0 (в случае статистической независимости $x^{(1)}$ и $x^{(2)}$)

до $+\infty$, и в условиях справедливости гипотезы о статистической независимости анализируемых признаков она приблизительно (асимптотически по $n \rightarrow \infty$) подчиняется з.р.в. χ^2 с $(m_1 - 1)(m_2 - 1)$ степенями свободы (для достижения удовлетворительной точности последнего утверждения требуется заполненность всех клеток таблицы сопряженности 11.2 ненулевыми элементами, удовлетворяющими условию $\min_{i,j} n_{ij} > 3$).

Таким образом, если мы хотим проверить гипотезу о статистически значимом отличии от нуля характеристики с.т.с.с. Y^2 (т.е. гипотезу о наличии связи между $x^{(1)}$ и $x^{(2)}$), то необходимо убедиться в выполнении неравенства

$$Y^2 > \chi_{\alpha}^2((m_1 - 1)(m_2 - 1)), \quad (11.64)$$

где $\chi_{\alpha}^2(m)$, как всегда, — 100 α %-ная точка χ^2 -распределения с m степенями свободы (см. соответствующую таблицу в Приложении 1), а α — заданный уровень значимости критерия, т.е. вероятность принять решение о наличии статистической связи между анализируемыми переменными в то время, как в действительности они являются статистически независимыми.

Если с помощью той или иной характеристики (X^2 или Y^2) мы убедились в том, что статистически значимая связь между $x^{(1)}$ и $x^{(2)}$ действительно существует, то возникает вопрос *интервальной оценки* характеристики этой связи. Для приближенного построения такой оценки мы должны уметь вычислить дисперсию соответствующей точечной оценки (см. п. 7.5.4), т.е. в нашем случае — дисперсии статистик X^2 и Y^2 в условиях, когда оцениваемый с их помощью теоретический показатель связи отличен от нуля.

Для характеристики X^2 и основанном на ней коэффициенте Крамера C существуют приближенные формулы¹:

$$\sigma_{X^2}^2 = DX^2 \approx 4 \left[X^2 + \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \left(\frac{\frac{n_{ij}}{n} - \frac{n_{i.}}{n} \frac{n_{.j}}{n}}{\frac{n_{i.}}{n} \frac{n_{.j}}{n}} \right)^2 - \frac{(X^2)^2}{n} \right], \quad (11.65)$$

$$\sigma_C^2 = DC \approx \frac{1}{n \min(m_1 - 1, m_2 - 1)}. \quad (11.66)$$

Заметим, что в формуле (11.65) первое слагаемое в квадратных скобках (т.е. X^2) является *главным членом*, так что два других слагаемых

¹ Точные формулы (они получены Холдуйном в 1939 г.) очень сложны, поэтому здесь приводится только приближенный вариант, полученный К. Пирсоном в 1915 г.

играют роль *остаточного члена*. Поэтому другой, более грубый (менее точный) вариант этой формулы имеет вид

$$\sigma_{X^2}^2 = DX^2 \approx 4X^2. \quad (11.65')$$

Соответственно в качестве приближенных доверительных интервалов для истинных значений коэффициента квадратичной сопряженности и коэффициента Крамера используются (при доверительной вероятности $P = 1 - 2\alpha$):

$$[X^2 - u_{1-\alpha}\sigma_{X^2}; \quad X^2 + u_{1-\alpha}\sigma_{X^2}] \quad (11.67)$$

и

$$[C - u_{1-\alpha}\sigma_C; \quad C + u_{1-\alpha}\sigma_C], \quad (11.68)$$

где u_q , как обычно, q -квантиль стандартного нормального распределения (см. соответствующую таблицу в приложении 1), а вместо σ_{X^2} и σ_C подставляются их приближенные значения, вычисленные соответственно по формулам (11.65) (или (11.65')) и (1.66).

Пример 11.9. В табл. 11.3 (из [Gilby W. H., Biometrika, 8, 94]) дано распределение 1725 школьников (т. е. $n = 1725$) по значениям двух анализируемых категоризованных признаков: по качеству одежды ($x^{(1)}$) и по их умственным способностям ($x^{(2)}$). Были определены 4 градации (т. е. $m_1 = 4$) по первому признаку (1: «очень хорошо»; 2: «хорошо»; 3: «удовлетворительно»; 4: «плохо») и 6 градаций (т. е. $m_2 = 6$) по второму признаку (1: «плохие»; 2: «ниже средних»; 3: «средние»; 4: «выше средних»; 5: «высокие»; 6: «превосходные»).

Таблица 11.3

Градации признака $x^{(1)}$	Градации признака $x^{(2)}$						Σ
	1	2	3	4	5	6	
1	33	48	113	209	194	39	636
2	41	100	202	255	138	15	751
3	39	58	70	61	33	4	265
4	17	13	22	10	10	1	73
Σ	130	219	407	535	375	59	1725

Нас интересует, существует ли связь между манерой одеваться и способностями, и если «да», то какова степень тесноты этой связи.

Вычисление величины X^2 по формуле (11.60) дает $X^2 = 174,92$ и соответственно значение коэффициента Крамера (см. формулу (11.62)) $C = 0,184$.

Поскольку даже при уровне значимости $\alpha = 0,001$ соответствующее пороговое значение $\chi^2_{0,001}(15) = 37,697$ оказывается намного меньшим статистики X^2 , то связь есть и, по-видимому, характеризуется *достаточно высокой степенью тесноты*. Если сопоставить этот вывод с выборочным значением коэффициента Крамера ($C = 0,184$), то мы получаем основание относиться к этой шкале иначе, чем, например, к шкале абсолютных значений парного коэффициента корреляции $|\hat{r}|$: если для $|\hat{r}|$ величина 0,184 означает отсутствие или крайне слабую связь между $x^{(1)}$ и $x^{(2)}$, то для C это значение, как мы видим, свидетельствует о наличии достаточно тесной связи.

Величина σ_C , вычисленная по приближенной формуле (11.66), оказывается равной $\sqrt{1/3n} = 0,014$, так что в соответствии с (11.68) можем сделать вывод, что истинное (теоретическое) значение коэффициента Крамера заключено в пределах от $0,184 - 1,96 \cdot 0,014 = 0,157$ до $0,184 + 1,96 \cdot 0,014 = 0,211$.

ВЫВОДЫ

1. Трудности, стоящие на пути практического использования в статистическом исследовании моделей *многомерных* з.р.в., обусловили тот факт, что в подавляющем большинстве случаев все выводы многомерного статистического анализа строятся лишь на базе оценок вектора средних значений и ковариационной матрицы исследуемого векторного признака.

2. Корреляционный анализ составляет содержание начальных этапов исследования всех трех центральных проблем многомерного статистического анализа. В *проблеме статистического исследования зависимостей* корреляционный анализ позволяет выявлять сам факт существования статистических связей между компонентами анализируемого векторного признака и оценивать степень тесноты этих связей. Для *проблем классификации объектов и признаков и снижения размерности анализируемого признакового пространства* он предлагает и оценивает подходящие характеристики парных отношений (ковариации, корреляции, разные виды парных сравнений), которые используются в дальнейшем в качестве базовой исходной информации.

3. Универсальной характеристикой степени тесноты статистической связи (с.т.с.) между результирующим *количественным* показателем y и

объясняющими количественными переменными $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^T$ является коэффициент детерминации y по X — $K_d(y; X)$. Другие распространенные характеристики с.т.с.с. (парные, частные и множественные коэффициенты корреляции, корреляционное отношение) представляют собой те или иные частные версии коэффициента детерминации, реализованные в рамках различных конкретных схем зависимостей.

4. *Парные корреляционные характеристики* позволяют измерять степень тесноты статистической связи между парой переменных *без учета опосредованного или совместного влияния других показателей*, вычисляются (оцениваются) они по результатам наблюдений только анализируемой пары показателей.

5. Факт установления тесной статистической связи между переменными не является, вообще говоря, достаточным основанием для доказательства существования *причинно-следственной связи между этими переменными*.

6. Парные и частные *коэффициенты корреляции* являются измерителями степени тесноты *линейной* связи между переменными. В этом случае корреляционные характеристики могут оказаться как положительными, так и отрицательными в зависимости от одинаковой или противоположной тенденции взаимосвязанного изменения анализируемых переменных. При положительных значениях коэффициента корреляции говорят о наличии *положительной линейной статистической связи*, при отрицательных — об *отрицательной*.

7. При наложении случайных ошибок на значения исследуемой пары переменных (например, ошибок измерения) оценка статистической связи между исходными переменными, построенная по наблюдениям, оказывается искаженной. В частности, получаемые при этом оценки коэффициентов корреляции будут *заниженными*. Существуют методы, позволяющие учесть это искажение.

8. Измерителем степени тесноты связи *любой формы* является *корреляционное отношение*, для вычисления которого необходимо разбить область значений предсказывающей переменной X на интервалы (гиперпараллелепипеды) группирования. Возможна параметрическая модификация корреляционного отношения, при которой вычисление соответствующих выборочных значений не требует предварительного разбиения на интервалы группирования.

9. *Частный коэффициент корреляции* позволяет оценить степень тесноты линейной связи между двумя переменными, *очищенной от опосредованного влияния* других факторов. Для его расчета необходима исходная информация как по анализируемой паре переменных, так и по всем тем

переменным, опосредованное («мешающее») влияние которых мы хотим элиминировать.

10. Анализ статистических связей между порядковыми переменными сводится к статистическому анализу различных упорядочений (ранжировок) одного и того же конечного множества объектов и осуществляется с помощью методов *ранговой корреляции*. В зависимости от типа изучаемой ситуации (шкала измерения анализируемого свойства не известна исследователю или отсутствует вовсе; существуют косвенные или частные количественные показатели, в соответствии со значениями которых можно определить место каждого объекта в общем ряду всех объектов, упорядоченных по анализируемому основному свойству) процесс упорядочения объектов производится либо с привлечением экспертов, либо формализованно — с помощью перехода от исходного ряда наблюдений косвенного количественного признака к соответствующему вариационному ряду.

11. Исходные статистические данные для проведения рангового корреляционного анализа представлены *таблицей (матрицей) рангов* статистически обследованных объектов размера $n \times (p + 1)$ (число объектов на число анализируемых переменных). При формировании матрицы рангов допускаются случаи неразличимости двух или нескольких объектов по изучаемому свойству («объединенные» ранги).

12. К *основным задачам* теории и практики ранговой корреляции относятся: анализ структуры исследуемой совокупности упорядочений (задача *A*); анализ интегральной (совокупной) согласованности рассматриваемых переменных и их условная ранжировка по критерию степени тесноты связи каждой из них со всеми остальными переменными (задача *B*); построение единого группового упорядочения объектов на основе имеющейся совокупности согласованных упорядочений (задача *C*).

13. В качестве основных характеристик парной статистической связи между упорядочениями используются *ранговые коэффициенты корреляции Спирмэна $\tau^{(S)}$ и Кендалла $\tau^{(K)}$* . Значения этих коэффициентов меняются в диапазоне от -1 до $+1$, причем экстремальные значения характеризуют связь соответственно пары прямо противоположных и пары совпадающих упорядочений, а нулевое значение рангового коэффициента корреляции получается при полном отсутствии статистической связи между анализируемыми порядковыми переменными.

14. В качестве основной характеристики статистической связи между *несколькими (m) порядковыми переменными* используется так называемый *коэффициент конкордации (согласованности) Кендалла $W(m)$* . Между значением этого коэффициента и значениями парных ранговых коэффициентов Спирмэна, построенных для каждой пары анализируемых

переменных, существуют простые соотношения.

15. Если представить себе, что каждому объекту некоторой достаточно большой гипотетической совокупности (будем называть ее *генеральной совокупностью*) приписан какой-то ранг по каждой из рассматриваемых переменных и что статистическому обследованию подлежит *лишь часть* этих объектов (*выборка объема n*), то достоверность и практическая ценность выводов, основанных на анализе ранговой корреляции, существенно зависят от ответа на вопрос: как ведут себя выборочные значения интересующих нас ранговых корреляционных характеристик при повторениях выборок заданного объема, извлеченных из этой генеральной совокупности. Это и составляет *предмет исследования статистических свойств выборочных ранговых характеристик связи*. Результаты этого исследования относятся прежде всего к построению правил проверки статистической значимости анализируемой связи и к построению доверительных интервалов для неизвестных значений коэффициентов связи, характеризующих всю генеральную совокупность.

16. Основными измерителями степени тесноты парной статистической связи между *категоризованными* переменными являются *коэффициент квадратичной сопряженности X^2* и *информационная характеристика связи Y^2* . Эти характеристики используют в ситуациях, когда анализируемыми переменными являются ординальные или номинальные признаки, шкала возможных «значений» которых определена заданным набором их состояний (*градаций*).

ГЛАВА 12. РАСПОЗНАВАНИЕ ОБРАЗОВ И ТИПОЛОГИЯ ОБЪЕКТОВ В СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ИССЛЕДОВАНИЯХ (МЕТОДЫ КЛАССИФИКАЦИИ)

12.1. Сущность, типологизация и прикладная направленность задач классификации объектов

Необходимость анализа и формализации задач, связанных со сравнением и классификацией объектов, сознавали ученые далекого прошлого. «Его (Аристотеля) величайшим и в то же время чреватым наиболее опасными последствиями вкладом в науку была идея классификации, которая проходит через все его работы ... Аристотель ввел или, по крайней мере, кодифицировал способ классификации предметов, основанный на сходстве и различии ...», — писал Дж. Берналл в «Науке истории общества» (М.: Изд-во иностр. лит., 1956, с. 117). После Аристотеля с его «деревом вещей жизни» имеется еще в докомпьютерной эре ряд интереснейших примеров прекрасно построенных классификаций как в естественных, так и в общественных науках. Упомянем здесь (в хронологическом порядке) две из них: а) иерархическая классификация (основанная на понятии сходства) растений и видов М. Адансона (1757 г.); б) знаменитая периодическая система элементов Д. И. Менделеева (1869 г.), представляющая собой по существу классификацию многомерных наблюдений (каждый химический элемент может быть представлен в виде вектора характеризующих его разнотипных признаков, включая характеристики конфигурации внешних электронных оболочек атомов) с выявленным единым классифицирующим фактором (зарядом атомного ядра) и с упорядочением элементов внутри каждого класса.

Однако до разработки аппарата многомерного статистического анализа и, главное, до появления и развития достаточно мощной электронно-

вычислительной базы проблемы теории и практики классификации относились не к разработке методов и алгоритмов, а к полноте и тщательности отбора и теоретического анализа изучаемых объектов, характеризующих их признаков, смысла и числа градаций по каждому из этих признаков.

Все методы классификации сводились по существу к методу так называемой *комбинационной группировки*, когда все характеризующие объект признаки носят дискретный характер или сводятся к таковым (пол или мотив миграции индивидуума, уровень жилищных условий или число детей в семье и т. п.), а *два объекта относятся к одной группе только при точном совпадении зарегистрированных на них градаций одновременно по всем характеризующим их признакам* (одинаковый пол, мотив миграции и т. д.).

Однако по мере роста объемов перерабатываемой информации и, в частности, числа классифицируемых объектов и характеризующих их признаков возможность эффективной реализации подобной логики исследования становилась все менее реальной (так, например, число k групп или классов, подсчитываемое при комбинационной группировке по формуле $k = m_1 \cdot m_2 \dots m_p$, где m_j — число градаций по признаку $x^{(j)}$, а p — общее число анализируемых признаков, уже при $m_j = 3$ и $p = 5$ оказывается равным 243). Именно электронно-вычислительная техника стала тем главным инструментом, который позволил по-новому подойти к решению этой важной проблемы и, в частности, конструктивно воспользоваться разработанным к этому времени мощным аппаратом многомерного статистического анализа: методами распознавания образов «с учителем» (дискриминантный анализ) и «без учителя» (автоматическая классификация, или кластер-анализ).

Развитие электронно-вычислительной техники как средства обработки больших массивов данных стимулировало проведение в последние годы широких комплексных исследований сложных социально-экономических, технических, медицинских и других процессов и систем, таких, как образ и уровень жизни населения, совершенствование организационных систем, региональная дифференциация социально-экономического развития, планирование и прогнозирование отраслевых систем, закономерности возникновения сбоев (в технике) или заболеваний (в медицине) и т. п. В связи с многоплановостью и сложностью этих объектов и процессов данные о них по необходимости носят *многомерный* и *разнотипный* характер, так как до их анализа обычно бывает неясно, насколько существенно то или иное свойство для конкретной цели. В этих условиях выходят на первый план проблемы построения группировок и классификаций по многомерным данным (т. е. проблемы *классификации многомерных наблюдений*),

причем появляется возможность оптимизации этого построения с точки зрения наибольшего соответствия получаемого результата поставленной конечной цели классификации.

Цели классификации существенно расширяются, и одновременно содержание самого процесса классификации становится неизмеримо богаче и сложнее. Оно, в частности, дополняется *проблемой построения самой процедуры классификации*, ранее носившей чисто технический характер.

Прежде чем переходить к примерам и типологизации (в прикладном и математическом аспектах) задач классификации, определим сам термин *классификация*.

В самой общей формулировке под *классификацией* мы будем понимать *разделение рассматриваемой совокупности объектов или явлений на однородные, в определенном смысле, группы либо отнесение каждого из заданного множества объектов к одному из заранее известных классов* (при этом классифицируемое «заданное множество» может состоять из единственного объекта). Заметим, что термин «классификация» используется, в зависимости от контекста, для обозначения как самого процесса «разделения–отнесения», так и его результата.

Сделаем шаг в сторону формализации общей задачи классификации и сформулируем ее теперь в терминах *статистического* варианта различных форм задания исходных статистических данных (см. (9.1) и (9.2) при $N = 1$) по схеме «на входе–на выходе задачи».

На «входе» задачи исследователь имеет:

(а) *n классифицируемых объектов*, представленных данными вида (9.1) (тогда каждая i -я строка матрицы (9.1) отражает значения p характеризующих i -й объект признаков $x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)}$) или данными вида (9.2) (тогда каждая i -я строка матрицы (9.2) задает попарные отношения $\gamma_{i1}, \gamma_{i2}, \dots, \gamma_{in}$ i -го объекта со всеми остальными классифицируемыми объектами);

(б) *обучающие выборки*

$$X_{j1}, X_{j2}, \dots, X_{jn_j}, \quad j = 1, 2, \dots, k, \quad (12.1)$$

каждая (j -я) из которых определяет значения анализируемых признаков $X_{ji} = (x_{ji}^{(1)}, x_{ji}^{(2)}, \dots, x_{ji}^{(p)})^T$ на n_j объектах (т.е. $i = 1, 2, \dots, n_j$), о которых *априори* известно, что все они принадлежат j -му классу, причем число k различных выборок (12.1) равно общему числу всех возможных классов (так что *каждый класс представлен своей порцией выборочных данных*); если классифицируемые данные представлены в форме (9.2), то каждая (j -я) из обучающих выборок также представляется матрицей вида

(9.2) размерности $n_j \times n_j$, однако этот вариант задачи «классификации с обучением» в книге не рассматривается.

На «выходе» задачи мы должны иметь результат в одной из двух форм:

(i) *если число классов k и их смысл известны заранее*, то каждое из n классифицируемых многомерных наблюдений должно быть снабжено «адресом» (номером) класса, к которому оно принадлежит;

(ii) *если число классов u (или) их смысл выявляются в процессе классификации*, то результатом классификации является разделение множества классифицируемых объектов на определенное число однородных (в определенном смысле) групп, каждая из которых объявляется «классом».

Если исследователь располагает на «входе» задачи не только классифицируемыми данными (а), но и обучающими выборками (б) (см. (12.1)), то говорят, что решается *задача классификации при наличии обучающих выборок* («классификация с обучением»); в противном случае речь идет о задаче «классификации без обучения».

Для пояснения сущности основных прикладных типов задач классификации и конечных прикладных целей, которые ставит при этом перед собой исследователь, рассмотрим примеры.

Пример 12.1. Выявление типологии потребительского поведения населения, анализ сущности дифференциации этого поведения, прогноз структуры потребления.

В качестве исходной информационной базы используются данные бюджетных обследований семей. Поясним логическую схему исследования. Многомерная статистика рассматривает совокупность изучаемых многомерных объектов как совокупность точек или векторов в пространстве описывающих их признаков. Применительно к схеме потребления совокупностью объектов, подлежащих изучению, является множество элементарных потребительских ячеек — семей. Каждая семья характеризуется, с одной стороны, некоторым набором X *факторов-детерминантов* (социально-демографические и другие признаки, описывающие условия жизнедеятельности семьи), а с другой — набором Y параметров поведения («переменных поведения»), в которых отражаются ее фактические потребности.

В качестве социально-демографических факторов, имеющих существенное значение для изучения потребительских аспектов социальной жизни, целесообразно использовать, например, общественную и национальную принадлежность, уровень образования и квалификацию, характер труда, демографический тип и возраст семьи, тип населенного пункта и характер жилища, размер и структуру имущества, уровень доходов.

Различия в потребностях, складывающиеся под влиянием социально-демографических и природно-климатических условий, являются объективно существующими; они формируют весь строй поведения потребителя в конкретно-исторических условиях, а в конечном счете порождают своеобразные типы потребителей, ориентированные на существенно разное потребление.

Весь комплекс социально-демографических и других факторов, существенно воздействующих на структуру потребления, будем называть *типообразующим*. Они имеют определяющее значение, в то время как все другие дают лишь случайную вариацию в пределах одной группы (типа) потребительского поведения.

В качестве признаков поведения Y можно рассматривать три группы параметров: а) уровень и структуру потребления; б) характер (объем и содержание) использования свободного времени; в) интенсивность изменения социального, трудового, демографического статуса.

Итак, рассматриваются числовые характеристики и градации типобразующих и одновременно поведенческих признаков каждой семьи из анализируемой совокупности.

Решение общей проблемы, связанной с выявлением и прогнозом структуры и дифференциации потребностей населения, распадается на следующие этапы.

1. *Сбор и первичная статистическая обработка исходных данных.* Исследуемые объекты (семьи) выступают в качестве многомерных наблюдений или точек в двух многомерных пространствах признаков. Фиксируя в качестве координат этих точек значения (или градации) типобразующих переменных X (т.е. факторов-детерминантов), рассматриваем их в «пространстве состояния» $\Pi(X)$ — пространстве, координатами которого служат основные показатели жизнедеятельности семей. Фиксируя же в качестве координат тех же самых объектов значения показателей Y их потребительского поведения, рассматриваем их в «пространстве поведения» $\Pi(Y)$. Очевидно, *при надлежащем выборе метрики в пространствах $\Pi(X)$ и $\Pi(Y)$ геометрическая близость двух точек в $\Pi(X)$ будет означать сходство условий жизнедеятельности соответствующих двух семей, так же как и геометрическая близость точек в $\Pi(Y)$ будет означать сходство их потребительского поведения.* Среди методов первичной статистической обработки анализируемых данных, обычно используемых на этой стадии исследования, широко распространенными и весьма полезными являются методы изучения различных одно-, дву- и трехмерных эмпирических распределений, которые сводятся к построению и различным представлениям (графическим, табличным) упомянутых вы-

ше комбинационных группировок. Пример табличного представления одной из таких двумерных комбинационных группировок приведен в табл. 12.1.

Эта комбинационная группировка построена на основе статистического обследования 400 семей по двум признакам из пространства $\Pi(X)$: по $x^{(1)}$ (руб.) — величине среднедушевого семейного дохода (с тремя градациями: «низкий», «средний» и «высокий»), и по $x^{(2)}$ — качеству жилищных условий (с четырьмя градациями: «низкое», «удовлетворительное», «хорошее» и «очень хорошее»). Каждая клетка таблицы соответствует классу, полученному в результате проведенной комбинационной группировки; внутри клетки обозначено число семей, имеющих данное сочетание градаций анализируемых признаков (подобные таблицы называют также «таблицами сопряженности», см. п. 11.4).

Таблица 12.1

Градация признака $x^{(1)}$ (доход)	Градация признака $x^{(2)}$ (жилищных условий)				Сумма
	«Низкое»	«Удовлетворительное»	«Хорошее»	«Очень хорошее»	
«Низкий»	24	12	4	0	40
«Средний»	20	100	140	20	280
«Высокий»	4	8	28	40	80
Сумма	48	120	172	60	400

Для более полного представления результатов подобной классификации можно было бы ввести в программу компьютера требование выпечатывать номера семей, попавших в каждую из двадцати клеточек таблицы.

Заметим, что непрерывным аналогом комбинационной группировки является обычный переход от исходных наблюдений непрерывной случайной величины к «группированным» выборочным (см. п. 6.1). Результат такого перехода представляется либо в виде таблицы, подобной табл. 12.1, либо в виде графика (гистограммы).

2. *Выявление основных типов потребления с помощью разбиения исследуемого множества точек-семей на классы в «пространстве поведения» $\Pi(Y)$.* Гипотеза существования «естественных», объективно обусловленных типов поведения, т. е. какого-то небольшого количества классов семей, таких, что семьи одного класса характеризуются сравнительно сходным, однотипным потребительским поведением, геометрически означает распадение исследуемой в «пространстве поведения» совокупности точек-семей на соответствующее число «сгустков» или «скоплений» то-

чек. Выявив с помощью подходящих методов многомерного статистического анализа (кластер-анализа, таксономии) эти классы-сгустки, тем самым определим основные типы потребительского поведения.

3. *Отбор наиболее информативных типобразующих признаков (факторов-детерминантов) и выбор метрики в пространстве типобразующих признаков.* Очевидно, неправомерно рассчитывать на то, что диапазоны возможных значений каждого из кандидатов в типобразующие признаки окажутся непересекающимися для семей с разным типом потребительского поведения. Другими словами, значения каждого из признаков $x^{(j)}$ в отдельности и их набора в совокупности подвержены некоторому неконтролируемому разбросу при анализе семей внутри каждого из типов потребления. Естественно считать наиболее информативными те факторы-детерминанты или те их наборы, разница в законах распределения которых оказывается наибольшей при переходе от одного класса потребительского поведения к другому. Эта идея и положена в основу метода отбора наиболее информативных (типобразующих) признаков-детерминантов.

4. *Анализ динамики структуры исследуемой совокупности семей в пространстве наиболее информативных типобразующих признаков.* Конечной целью этого этапа является прогноз тех постепенных преобразований классификационной структуры совокупности потребителей (семей, рассматриваемых в пространстве типобразующих признаков), которые должны произойти с течением времени.

5. *Прогноз структуры потребления.* На этом этапе исследования опираемся на результаты, полученные в итоге проведения предыдущего этапа, т. е. исходим из заданной классификационной структуры потребителей в интересующий нас период времени в будущем. Восстанавливая классификационную структуру потребления (классификационную структуру совокупности семей в пространстве признаков $\Pi(Y)$, характеризующих потребительское поведение семьи) по классификационной структуре потребителей (по классификационной структуре той же совокупности, но в пространстве типобразующих признаков), будем относить каждую конкретную семью к тому типу потребления, для которого значения характеризующих ее типобразующих признаков являются, грубо говоря, наиболее типичными.

Пример 12.2. Классификация как необходимый предварительный этап статистической обработки многомерных данных. Пусть исследуется зависимость интенсивности миграции населения $x^{(p)}$ (профессиональной или территориальной) от ряда социально-экономических и географических факторов $x^{(1)}, x^{(2)}, \dots, x^{(p-1)}$, таких, как

средний заработок, обеспеченность жилой площадью, детскими учреждениями, уровень образования, возможности профессионального роста, климатические условия и т.п. Естественно предположить (и результаты исследования это подтверждают), что для различных однородных групп индивидуумов одни и те же факторы влияют на $x^{(p)}$ в разной степени, а иногда и в противоположных направлениях. Поэтому до применения аппарата регрессионно-корреляционного анализа следует разбить все имеющиеся в нашем распоряжении данные $X_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)})^T$ ($i = 1, 2, \dots, n$) на однородные классы и *решать далее поставленную задачу отдельно для каждого такого класса*. Только в этом случае можно ожидать, что полученные коэффициенты регрессии $x^{(p)}$ по $x^{(1)}, x^{(2)}, \dots, x^{(p-1)}$ будут допускать содержательную интерпретацию, а мера тесноты связи между $x^{(p)}$ и $x^{(1)}, x^{(2)}, \dots, x^{(p-1)}$ окажется достаточно высокой.

Другой вариант такого рода примера получим, если в качестве объектов исследования рассмотрим предприятия определенной отрасли, а в качестве вектора наблюдений X_i — совокупность объективных (нерегулируемых) условий работы i -го обследованного предприятия (сырье, энергия, оснащенность техникой и рабочей силой и т.п.). Классификация предприятий по X производится как необходимый предварительный этап для возможности последующей объективной оценки работы коллективов и разработки обоснованных дифференцированных нормативов: очевидно, лишь к предприятиям, попавшим в один класс по X , может быть применена одинаковая система нормативов и стимулирующих показателей. Далее можно рассматривать задачу, аналогичную сформулированной выше, а именно: $Y = (y^{(1)}, \dots, y^{(q)})^T$ — вектор показателей качества работы предприятия (объем и качество выпускаемой продукции, ее себестоимость, рентабельность и т.п.), а $U = (u^{(1)}, \dots, u^{(m)})^T$ — вектор регулируемых факторов, от которых зависят условия производства (число основных подразделений, уровень автоматизации и т.д.), то задачу описания интересующей нас зависимости вида $Y = f(U)$ естественно решать отдельно для каждого однородного по X класса.

Пример 12.3. Классификация в задачах планирования выборочных обследований. Здесь речь пойдет о планировании выборочных экономико-социологических обследований городов. Предположим, что необходимо достаточно детально проанализировать подробные статистические данные о городах с целью выявления наиболее характерных черт в экономико-социологическом облике типичного среднерусского города. Производить подробный, кропотливый анализ по каждому из городов Российской Федерации, очевидно, слишком трудоемко, да и нецелесообразно. По-видимому, разумнее попытаться предварительно выявить число и со-

став различных типов в совокупности обследованных городов по набору достаточно агрегированных признаков $x^{(1)}, x^{(2)}, \dots, x^{(p-1)}$, характеризующих каждый город (например, понимать под $x^{(j)}$ число жителей города, приходящееся на каждую тысячу жителей, обладающих заданным j -м признаком, скажем, высшим образованием, специальностью металлурга и т.п.). А затем, отметив наиболее типичные города в каждом классе (наблюдения-точки X_i , наиболее близко располагающиеся к «центрам тяжести» своих классов), отобрать их для дальнейшего (более детализированного) социально-экономического анализа. При этом, очевидно, мера представительности отобранных «типичных городов» определится удельным весом количественного состава точек данного класса среди всех рассматриваемых точек (городов).

Анализ рассмотренных примеров с учетом, конечно, и другого накопившегося к настоящему времени опыта решения практических задач классификации в экономике, социологии, психологии и других сферах практической и научной деятельности человека позволяет произвести определенную систематизацию этих задач в соответствии с конечными прикладными целями исследования (табл. 12.2).

Таблица 12.2

№ п/п	Тип задачи классификации	Варианты (примеры) конечных прикладных целей исследования для данного типа задачи классификации
1	2	3
1	<i>Комбинационные группировки и их непрерывные обобщения</i>	1.1. Составление частотных таблиц и графиков, характеризующих распределение статистически обследованных объектов по градациям или интервалам группирования характеризующих их признаков (см. п. 1 в примере 12.1)
2	<i>Простая типологизация: выявление «стратификационной структуры» множества статистически обследованных объектов, «нащупывание» и описание четко выраженных скоплений («сгустков», «клас-</i>	2.1. Классификация как необходимый предварительный этап исследования, когда до проведения основной статистической обработки множества анализируемых данных (построения регрессионных моделей, оценки параметров генеральной совокупности и т.д.) до-

Продолжение таблицы 12.2

1	2	3
	теров», «образов», классов) этих объектов в анализируемом многомерном пространстве и построение правила отнесения каждого нового объекта к одному из выявленных классов	бываются расслоения этого множества на однородные (в смысле проводимого затем статистического анализа) группы данных (см. примеры 12.1 и 12.2). 2.2. Выявление и описание расслоенной природы анализируемой совокупности статистически обследованных объектов с целью формирования плана выборочных (например, экономико-социологических) обследований этой совокупности (см. пример 12.3). 2.3 Первый шаг в построении связанных типологий (см. п. 3 таблицы и пример 12.1)
3	<i>Связная неупорядоченная типологизация:</i> исследование зависимостей между не поддающимися упорядочению классификациями одного и того же множества объектов в разных признаковых пространствах, одно из которых построено на <i>результатирующих</i> (поведенческих) признаках (отражающих специфику функционирования объекта, его социально-экономическое поведение, состояние здоровья и т. п.), а другое — на <i>описательных</i> (отражающих условия функционирования и другие характеристики, от которых могут зависеть значения результирующих признаков)	3.1. Прогноз экономико-социологических ситуаций или отдельных социально-экономических показателей, включая задачу выявления так называемых типобразующих признаков, в том числе латентных, т. е. непосредственно не наблюдаемых (см. пример 12.1). 3.2. Диагностика в промышленности, технике, георазведке, медицине 3.3. Кластер-анализ, автоматическое (машинное) распознавание образов — зрительных, слуховых

Продолжение таблицы 12.2

1	2	3
4	Связная упорядоченная типологизация: модификация связной неупорядоченной типологизации (см. п. 3 таблицы), обусловленная дополнительным допущением, что классы, получаемые в пространстве результирующих (поведенческих) признаков, поддаются экспертному упорядочению по некоторому сводному (как правило, латентному, непосредственно не наблюдаемому) свойству: эффективности функционирования, качеству, степени прогрессивности (оптимальности) поведения и т. п.	4.1. Построение и интерпретация единого (свободного) латентного признака-классификатора в виде функции от исходных описательных признаков: классификация химических элементов по заряду их атомного ядра (периодическая система Д. И. Менделеева); построение фактора общей одаренности в педагогике и психологии; построение сводного показателя эффективности функционирования предприятия (см. п. 13.5); построение интегральной характеристики уровня мастерства спортсменов в игровых видах спорта (см. п. 13.5)
5	Структурная типологизация: дополнение и развитие простой типологизации (см. п. 2 таблицы) в направлении изучения и описания структуры взаимосвязей полученных классов, включая построение соответствующих иерархических систем (на классифицируемых элементах и на классах элементов), анализ роли и места каждого элемента и класса в общей структурной классификационной схеме. При этом структурная классификационная схема определяется составляющими ее	5.1. Классификация задач многоцелевого комплекса (крупной программы, научного направления производственного комплекса и т. п.) 5.2. Классификация элементов и подсистем по их функциональному назначению (производств — в территориально-производственном комплексе, территориальных единиц — в народнохозяйственном разделении труда и потребления, элементов организационных структур и т. п.). 5.3. Классификация лиц, принимающих решение, по их роли и близости позиций в понимании ситуации и способе решения задачи.

Продолжение таблицы 12.2

1	2	3
	классами (подсистемами) и характеристиками (правилами) их взаимодействия	5.4. Классификация исследуемых признаков и анализ структуры связей между ними
6	<i>Классификация динамических траекторий развития систем:</i> типологизация траекторий многомерных временных рядов $X(t) = (x^{(1)}(t), \dots, x^{(p)}(t))^T$, среди компонент $x^{(j)}(t)$ которых могут быть как количественные, так и качественные переменные	6.1. Задачи анализа типов динамики семейной структуры, потребительского поведения семей и др.

В качестве комментария к табл. 12.2 поясним методологическую общность задач 3.1–3.3: прогноза экономико-социологических ситуаций, диагностики и автоматического распознавания зрительных и слуховых образов. Для этого лежащую в основе их решения методологическую схему связной неупорядоченной типологизации представим следующим образом. Пусть в качестве исходных данных об объекте O_i ($i = 1, 2, \dots, n$) имеем вектор *описательных* (объясняющих) признаков $X_i = (x_i^{(1)}, \dots, x_i^{(p)})^T$ (это, в частности, характеристики условий жизнедеятельности i -й обследованной семьи в примере 12.1, значения параметров исследуемого технологического процесса или результаты обследований i -го пациента в задачах диагностики, геометрические или частотные характеристики распознаваемого образа в п. 3.3) и некоторую *информацию* Y_i *о том результирующем свойстве*, по которому производится классификация объектов (специфика социально-экономического поведения i -й семьи в примере 12.1).

Разница между задачами типа 3.1 и задачами 3.2 и 3.3 заключается в том, что в задачах прогноза экономико-социологических ситуаций информация Y_i об исследуемом результирующем свойстве объекта *не является окончательной*, т.е. не задает однозначно, как это делается в задачах 3.2 и 3.3, образа (класса, типа), к которому относится этот объект. Эта информация в задачах типа 3.1 носит лишь *промежуточный характер* и представляется, как правило, в виде вектора результирующих показателей $Y_i = (y_i^{(1)}, \dots, y_i^{(q)})^T$. Поэтому в отличие от задач 3.2 и 3.3 (в которых уже «на входе» задачи имеем распределение анализируе-

мых объектов-векторов X_i по классам, что и составляет так называемую «обучающую выборку») в задачах типа 3.1 нужно предварительно осуществить простую типологизацию множества объектов $\{O_i\} (i = 1, \dots, n)$ в пространстве результирующих показателей и лишь затем использовать полученные в результате этой типологизации классы в качестве обучающих выборок для построения классифицирующего правила в пространстве описательных признаков $\Pi(X)$.

«На выходе» же всех задач типа 3.1–3.3 должны быть 1) набор наиболее информативных объясняющих переменных (так называемых *типообразующих признаков*) $z^{(1)}(X), z^{(2)}(X), \dots, z^{(p')}(X)$, которые либо отбираются по определенному правилу из числа исходных описательных признаков $x^{(1)}, x^{(2)}, \dots, x^{(p-1)}$, либо строятся в качестве некоторых их комбинаций; 2) правило отнесения (*дискриминантная функция, классификатор*) каждого нового объекта O^* , заданного значениями своих описательных признаков X^* , к одному из заданных (или выявленных в процессе предварительной простой типологизации) в пространстве $\Pi(Y)$ классов или образов. При этом типобразующие признаки $Z = (z^{(1)}(X), \dots, z^{(p')}(X))^T$ и искомое правило классификации должны быть подобраны таким образом, чтобы обеспечивать наивысшую (в определенном смысле) точность решения задачи отнесения объекта к одному из анализируемых классов по заданным значениям его описательных признаков X .

Типологизация математических постановок задач классификации. Целесообразность и эффективность применения тех или иных методов классификации так же, как их предметная осмысленность, обусловлены конкретизацией базовой математической модели, т. е. математической постановкой задачи. Определяющим моментом в выборе математической постановки задачи является ответ на вопрос, *на какой априорной информации строится модель*. При этом априорная информация складывается из двух частей:

1) *из априорных сведений об исследуемых классах*; 2) *из априорной статистической (выборочной) информации*, т. е. так называемых *обучающих выборок* (их определения даны выше).

Априорные сведения об исследуемых генеральных совокупностях относятся обычно к виду или некоторым общим свойствам закона распределения исследуемого случайного вектора X в соответствующем пространстве и получаются либо из теоретических, предметно-профессиональных соображений о природе исследуемого объекта, либо как результат предварительных исследований. Получение априорной выборочной информации в экономике и социологии, как правило, связано с организацией систе-

мы экспертных оценок или с проведением специального предварительного этапа, посвященного решению задачи простой типологизации анализируемых объектов в пространстве результирующих показателей (см. пример 12.1).

Классификация задач разбиения объектов на однородные группы (в зависимости от наличия априорной и предварительной выборочной информации) и соответствующее распределение описания аппарата решения этих задач по пунктам данной главы представлены в табл. 12.3.

Таблица 12.3

Априорные сведения о классах (генеральных совокупностях)	Предварительная выборочная информация	
	Нет информации	Есть обучающие выборки
Некоторые самые общие предположения о законе распределения исследуемого вектора: гладкость, сосредоточенность внутри ограниченной области и т. п.	Классификация без обучения: кластер-анализ, таксономия, распознавание образов «без учителя», иерархические классификации (см. п. 12.4)	Непараметрические методы дискриминантного анализа (см. п. 12.2.3)
Различаемые генеральные совокупности заданы в виде параметрического семейства законов распределения вероятностей (параметры неизвестны)	Интерпретация исследуемой генеральной совокупности как смеси нескольких генеральных совокупностей. «Расщепление» этой смеси с помощью методов оценивания неизвестных параметров (см. п. 12.3)	Параметрические методы дискриминантного анализа (см. пп. 12.2.3, 12.2.4)
Различаемые генеральные совокупности заданы однозначным описанием соответствующих законов	Классификация при полностью описанных классах: различение статистических гипотез (см. гл. 8)	Обучающие выборки не нужны

12.2. Классификация при наличии обучающих выборок (дискриминантный анализ)

12.2.1. Класс как генеральная совокупность и базовая идея вероятностно-статистических методов классификации

Ставя задачу отнесения каждого из классифицируемых наблюдений $X_i = (x^{(1)}, x^{(2)}, \dots, x^{(p-1)})^T$, $i = 1, 2, \dots, n$, к одному из классов, необходимо четко определить понятие класса. Во всех постановках задач этого и следующего (п. 12.3) параграфов мы будем понимать под «классом» генеральную совокупность, описываемую одномодальной функцией плотности $f(X)$ (или одномодальным полигоном вероятностей в случае дискретных признаков X).

Для пояснения общей идеи, заложенной в основу построения всех вероятностно-статистических методов классификации, вернемся к примеру п. 7.5.1 и рис. 7.2. По существу уже в этом примере мы имеем дело с задачей отнесения трех наблюдений $y_1 = 5,16$, $y_2 = 5,25$ и $y_3 = 5,32$ к одной из двух гипотетических нормальных генеральных совокупностей (т. е. к одному из двух классов), различающихся между собой средними значениями. И решение мы приняли в пользу класса со средним значением $a_1 = 5,243$ потому, что в рамках этого класса наши наблюдения выглядят более естественными, более правдоподобными (что определяется произведением соответствующих им ординат плотности этого закона). Именно этот принцип и положен в основу вероятностных методов классификации: наблюдение будет относиться к тому классу (т. е. к той генеральной совокупности), в рамках которого (которой) оно выглядит более правдоподобным. Правда, во-первых, этот принцип может корректироваться с учетом удельных весов классов и специфики так называемой «функции потерь» $c(j | i)$, которая определяет стоимость потерь от отнесения объекта i -го класса к классу с номером j . И во-вторых, для того чтобы этот принцип практически реализовать, мы должны располагать полным описанием гипотетических классов, т. е. знанием функций $f_1(X), f_2(X), \dots, f_k(X)$, задающих з.р.в. соответственно для 1-го, 2-го, ..., k -го классов. Последнее затруднение обходят с помощью обучающих выборок в случае классификации с обучением, — и этому посвящен данный параграф, — и с помощью модели смеси распределений в случае классификации без обучения (этому посвящен п. 12.3).

12.2.2. Функции потерь и вероятности неправильной классификации

Очевидно, желательно строить такие методы классификации, которые минимизируют потери (или вероятность) неправильной классификации объектов. Посмотрим, как эти *две характеристики качества метода классификации* связаны между собой.

Мы уже ввели величину $c(j | i)$ потерь, которые мы несем при отнесении *одного* объекта i -го класса к классу j (при $i = j$, очевидно, $c_{ij} = 0$). Следовательно, если в процессе классификации мы поступили таким образом $m(j | i)$ раз, то потери, связанные с отнесением объектов i -го класса к классу j составят $m(j | i)c(j | i)$. Для того чтобы подсчитать общие потери C_n при такой процедуре классификации, надо просуммировать величину произведения $m(j | i)c(j | i)$ по всем $i = 1, 2, \dots, k$ и $j = 1, 2, \dots, k$, т. е.

$$C_n = \sum_{i=1}^k \sum_{j=1}^k c(j | i) m(j | i). \quad (12.2)$$

Для того чтобы потери не зависели от числа n классифицируемых объектов (а величина C_n , очевидно, будет расти с ростом n), перейдем к *удельной* характеристике потерь, разделив обе части (12.2) на n , а затем перейдем к пределу по $n \rightarrow \infty$:

$$\begin{aligned} C &= \lim_{n \rightarrow \infty} \left(\frac{1}{n} C_n \right) = \lim_{n \rightarrow \infty} \sum_{i=1}^k \sum_{j=1}^k c(j | i) \frac{m(j | i)}{n_i(n)} \frac{n_i(n)}{n} \\ &= \sum_{i=1}^k \pi_i \sum_{j=1}^k c(j | i) P(j | i). \end{aligned} \quad (12.3)$$

Предел в (12.3) понимается в смысле *сходимости по вероятности частот* $m(j | i)/n_i(n)$ и $n_i(n)/n$ соответственно к вероятностям $P(j | i)$, — отнести объект класса i к классу j , и π_i , — извлечения объекта класса i из общей совокупности анализируемых объектов; величину π_i называют также *априорной вероятностью* (или *удельным весом*) класса i (о сходимости по вероятности относительных частот см. п. 6.2).

Величина

$$C^{(i)} = \sum_{j=1}^k c(j | i) P(j | i) \quad (12.4)$$

определяет, как легко видеть, средние потери от неправильной классификации объектов i -го класса, так что *средние удельные потери* от непра-

вильной классификации всех анализируемых объектов будут:

$$C = \sum_{i=1}^k \pi_i C^{(i)}. \quad (12.3')$$

В достаточно широком классе ситуаций полагают, что потери $c(j | i)$ одинаковы для любой пары i и j , т.е.

$$c(j | i) = c_0 = \text{const} \quad \text{при} \quad j \neq i; \quad i, j = 1, 2, \dots, k.$$

В этом случае *стремление к минимизации средних удельных потерь C будет эквивалентно стремлению максимизации вероятности правильной классификации объектов, равной $\sum_{i=1}^k \pi_i P(i | i)$* . Действительно,

$$\begin{aligned} C &= \sum_{i=1}^k \pi_i \sum_{j=1}^k c(j | i) P(j | i) = c_0 \sum_{i=1}^k \pi_i \left(\sum_{\substack{j=1 \\ (j \neq i)}}^k P(j | i) \right) \\ &= c_0 \sum_{i=1}^k \pi_i (1 - P(i | i)) = c_0 \left(1 - \sum_{i=1}^k \pi_i P(i | i) \right) \end{aligned}$$

(при выводе этого соотношения мы воспользовались тем, что $c(i | i) = 0$ и $\sum_{j=1}^k P(j | i) = 1$ для любого i).

Поэтому часто при построении процедур классификации говорят не о потерях, а о *вероятностях неправильной классификации* $1 - \sum_{i=1}^k \pi_i P(i | i)$.

12.2.3. Принципиальное решение общей задачи построения оптимальных (байесовских) процедур классификации

Сформулируем постановку задачи построения оптимальной процедуры классификации p -мерных наблюдений

$$X_1, X_2, \dots, X_n \quad (12.5)$$

при наличии обучающих выборок (12.1).

Классифицируемые наблюдения (12.5) интерпретируются в данной задаче как выборка из генеральной совокупности, описываемой так называемой смесью k классов (одномодальных генеральных совокупностей)

с плотностью вероятности

$$f(X) = \sum_{j=1}^k \pi_j f_j(X), \quad (12.6)$$

где π_j — априорная вероятность появления в этой выборке элемента из класса (генеральной совокупности) j с плотностью $f_j(x)$ или, другими словами, π_j — это удельный вес элементов j -го класса в общей генеральной совокупности (12.6).

Введем понятие *процедуры классификации (решающего правила, дискриминантной функции)* $\delta(X)$. Функция $\delta(X)$ может принимать только целые положительные значения $1, 2, \dots, k$, причем те X , при которых она принимает значение, равное j , мы будем относить к классу j , т. е.

$$S_j = \{X: \delta(X) = j\}, \quad j = 1, 2, \dots, k.$$

Очевидно, S_j — это p -мерные области в пространстве $\Pi(X)$ возможных значений анализируемого многомерного признака X , причем функция $\delta(X)$ строится таким образом, чтобы их сумма (теоретико - множественная, см. п. 1.2) $S_1 + S_2 + \dots + S_k$ заполняла все пространство $\Pi(X)$ и чтобы они попарно не пересекались. Таким образом, решающее правило $\delta(X)$ может быть задано разбиением

$$S = (S_1, S_2, \dots, S_k) \quad (12.7)$$

всего пространства $\Pi(X)$ на k непересекающихся областей.

Процедура классификации (дискриминантная функция) $\delta(X)$ (или S) называется оптимальной (байесовской), если она сопровождается минимальными потерями (12.3) среди всех других процедур классификации.

Оказывается (см., например, [Андерсон Т.]), процедура классификации $S^{(\text{опт})} = (S_1^{(\text{опт})}, \dots, S_k^{(\text{опт})})$, при которой потери (12.3) будут оптимальными, определяется следующим образом:

$$S_j^{(\text{опт})} = \left\{ X: \sum_{\substack{i=1 \\ i \neq j}}^k \pi_i f_i(X) c(j | i) = \min_{1 \leq l \leq k} \sum_{\substack{i=1 \\ i \neq l}}^k \pi_i f_i(X) c(l | i) \right\}. \quad (12.8)$$

Другими словами, наблюдение X_ν ($\nu = 1, 2, \dots, n$) будет отнесено к классу j тогда, когда средние удельные потери от его отнесения именно в этот класс окажутся минимальными по сравнению с аналогичными потерями, связанными с отнесением этого наблюдения в любой другой класс.

Относительно простой вид приобретает правило классификации (12.8) в случае равных потерь $c(j | i)$ (т.е. при выполнении соотношения $c(j | i) = c_0 = \text{const}$). В этом случае наблюдение X_ν будет отнесено к классу j тогда, когда

$$\pi_j f_j(X_\nu) = \max_{1 \leq i \leq k} \pi_i f_i(X_\nu) \quad (12.8')$$

(т.е. максимизируется «взвешенная правдоподобность» этого наблюдения в рамках класса, где в качестве весов выступают априорные вероятности π_j).

Однако соотношения (12.8) и (12.8') задают нам лишь *теоретическое* оптимальное правило классификации: для того чтобы его реально построить, необходимо знание априорных вероятностей $\pi_1, \pi_2, \dots, \pi_k$ и з.р.в. $f_1(X), f_2(X), \dots, f_k(X)$. В статистическом варианте решения этой задачи данные величины заменяются соответствующими оценками, построенными на базе обучающих выборок (12.1).

Априорные вероятности π_j ($j = 1, 2, \dots, k$) оцениваются просто, если ряд наблюдений, составленный из всех обучающих выборок (12.1), может быть классифицирован как случайная выборка объема $n_{\text{об}} = n_1 + n_2 + \dots + n_k$ из генеральной совокупности (12.6). Тогда оценки

$$\hat{\pi}_j = \frac{n_j}{n_{\text{об}}}, \quad (12.9)$$

где n_j — объем j -й обучающей выборки.

Впрочем, величины π_j часто определяются априори самой содержательной сущностью задачи.

Что касается задачи оценки з.р.в. $f_1(X), \dots, f_k(X)$, то ее удобно разбить на два случая:

1-й случай (параметрический дискриминантный анализ) характеризуется известным общим видом функций $f_j(X)$, т.е. все классы описываются з.р.в. одного и того же параметрического семейства $\{f(X; \Theta)\}$: класс i отличается от класса j только значением параметра Θ , т.е.

$$f_j(X) = f(X; \Theta_j), \quad j = 1, 2, \dots, k. \quad (12.10)$$

Тогда в качестве оценок $\hat{f}_j(X)$ неизвестных функций $f_j(X)$ используются функции $\hat{f}_j(X; \hat{\Theta}_j)$, где $\hat{\Theta}_j$ — статистическая оценка неизвестного значения параметра Θ_j , полученная по наблюдениям j -й обучающей выборки (12.1). Пример реализации этой схемы приведен в следующем пункте.

2-й случай (*непараметрический дискриминантный анализ*) не предусматривает знания общего вида функций $f_j(X)$ ($j = 1, 2, \dots, k$). В этом случае приходится строить так называемые *непараметрические оценки* $\hat{f}_j(X)$ для функций $f_j(X)$, например, гистограммного или ядерного типа, либо пользоваться некоторыми специальными приемами (см., например, [Айвазян С. А., Бежаева З. И., Староверов О. В., с. 44–50]).

12.2.4. Параметрический дискриминантный анализ в случае нормальных классов

Предположим, что класс j идентифицируется p -мерной нормальной генеральной совокупностью с вектором средних значений $\hat{\mathbf{a}}_j$ и ковариационной матрицей Σ (общей для всех классов), $j = 1, 2, \dots, k$. Тогда при реализации общих оптимальных правил классификации (12.8) и (12.8') в качестве функций $f_j(X)$ следует использовать p -мерные нормальные плотности $\varphi(X; \hat{\mathbf{a}}_j, \hat{\Sigma})$ (см. (3.9')), где оценки для векторов средних значений $\hat{\mathbf{a}}_j = (\hat{a}_j^{(1)}, \hat{a}_j^{(2)}, \dots, \hat{a}_j^{(p)})^T$ и для ковариационной матрицы $\hat{\Sigma} = (\hat{\sigma}_{lq})$, $l, q = 1, 2, \dots, p$, получены с помощью метода максимального правдоподобия (см. п. 7.5.1) по обучающим выборкам (12.1); они имеют вид

$$\hat{a}_j^{(l)} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ji}^{(l)}, \quad l = 1, 2, \dots, p, \quad j = 1, 2, \dots, k; \quad (12.11)$$

$$\hat{\sigma}_{lq} = \frac{1}{n_{об} - k} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ji}^{(l)} - \hat{a}_j^{(l)})(x_{ji}^{(q)} - \hat{a}_j^{(q)}), \quad (12.12)$$

$$l, q = 1, 2, \dots, p,$$

$$n_{об} = n_1 + n_2 + \dots + n_k.$$

Правило классификации (12.8') наблюдений (12.5) имеет весьма простой вид в случае $k = 2$. Действительно, правило (12.8') эквивалентно следующему: наблюдение X_ν относится к классу с номером j_0 тогда и только тогда, когда

$$\frac{f_{j_0}(X_\nu)}{f_j(X_\nu)} \geq \frac{\pi_j}{\pi_{j_0}} \quad \text{для всех } j = 1, 2, \dots, k, \quad (12.13)$$

или, что то же,

$$\ln \frac{f_{j_0}(X_\nu)}{f_j(X_\nu)} \geq \ln \frac{\pi_j}{\pi_{j_0}} \quad \text{для всех } j = 1, 2, \dots, k. \quad (12.13')$$

В случае *нормальных* плотностей $f_j(X)$, т. е. функций $f_j(X)$, определяемых по формуле (3.9') с заменой M и Σ соответственно векторами $\hat{a}_j$ и матрицей $\hat{\Sigma}$, определяемыми по формулам (12.11)–(12.12), соотношение (12.13') эквивалентно соотношению

$$\left[X_\nu - \frac{1}{2}(\hat{a}_{j_0} + \hat{a}_j) \right]^\top \hat{\Sigma}^{-1}(\hat{a}_{j_0} - \hat{a}_j) \geq \ln \frac{\pi_j}{\pi_{j_0}} \quad (12.13'')$$

для всех $j = 1, 2, \dots, k$.

Соотношение (12.13''), таким образом, задает вид *дискриминантной функции в задаче различения нормальных классов при постоянных значениях потерь от неправильной классификации*. Особенно простым становится это правило классификации в случае двух классов ($k = 2$) и одинаковых априорных вероятностей ($\pi_1 = \pi_2 = 0,5$). В этом случае наблюдение X_ν следует относить к 1-му классу тогда и только тогда, когда

$$\left[X_\nu - \frac{1}{2}(\hat{a}_1 + \hat{a}_2) \right]^\top \hat{\Sigma}^{-1}(\hat{a}_1 - \hat{a}_2) \geq 0, \quad (12.13''')$$

и ко 2-му классу во всех остальных случаях. Легко понять, что при классификации на два класса *одномерных* ($p=1$) нормальных наблюдений решение об отношении наблюдения x_ν к одному из двух классов будет определяться знаком произведения $[x_\nu - (\hat{a}_1 + \hat{a}_2)/2](\hat{a}_1 - \hat{a}_2)$.

Пример 12.4. Специальное исследование показало, что склонность фирм к утаиванию части своих доходов (и, соответственно, — к уклонению от уплаты части налогов) в существенной мере определяется двумя показателями:

- $x^{(1)}$ — соотношением «быстрых активов» и текущих пассивов;
- $x^{(2)}$ — соотношением прибыли и просроченных платежей (оба показателя оцениваются по определенной методике в шкале от 300 до 900 баллов).

В табл. 12.4 представлены значения этих показателей (данные налоговой инспекции) по 10 фирмам, уличенным в тех или иных формах уклонения от уплаты налогов ($n_1 = 10$), и по 13 фирмам, не имеющим замечаний по уплате налогов ($n_2 = 13$). Кроме того, имеющаяся статистика и специальные обследования свидетельствуют о том, что доля фирм, в той или иной форме уклоняющихся от уплаты налогов, достигает 50% (т. е. $\pi_1 = \pi_2 = 0,5$). Наконец, статистическая проверка гипотез о нормальном характере распределения двумерного признака $X = (x^{(1)}, x^{(2)})^\top$ *внутри* каждой из анализируемых совокупностей фирм — уклоняющихся от уплаты налогов (совокупности 1) и платящих налоги (совокупности 2),

и о равенстве их ковариационных матриц $\Sigma(1)$ и $\Sigma(2)$ дала неотрицательный результат, т.е. можно считать, что имеющиеся у нас обучающие выборки (1) и (2) извлечены из *нормальных* генеральных совокупностей с *одинаковыми* ковариационными матрицами.

Таблица 12.4

№ № пп (i)	Обучающая выборка (1) (фирмы, уклоняющиеся от налогов)		Обучающая выборка (2) (фирмы, не уклоняю- щиеся от налогов)	
	$x_{1i}^{(1)}$	$x_{1i}^{(2)}$	$x_{2i}^{(1)}$	$x_{2i}^{(2)}$
1	740	680	750	590
2	670	600	360	600
3	560	550	720	750
4	540	520	540	710
5	590	540	570	700
6	590	700	520	670
7	470	600	590	790
8	560	540	670	700
9	540	630	620	730
10	500	600	690	840
11	-	-	610	680
12	-	-	550	730
13	-	-	590	750
среднее	$\bar{a}_1^{(1)} = 576,0$	$\bar{a}_1^{(2)} = 596,0$	$\bar{a}_2^{(1)} = 598,5$	$\bar{a}_2^{(2)} = 710,8$

На фирме, не прошедшей проверку налоговой инспекции, зарегистрированы значения переменных $X = (x^{(1)}, x^{(2)})^T$: $x_0^{(1)} = 740$, $x_0^{(2)} = 590$. *Определить, к какой совокупности (1 или 2) будет отнесена эта фирма (т.е. наблюдение $X_0 = (x_0^{(1)}, x_0^{(2)})^T$), если воспользоваться методом параметрического дискриминантного анализа (с учетом нормальности анализируемых наблюдений).*

Из условия задачи следует, что мы должны воспользоваться правилом классификации, представленным соотношением (12.13'').

Необходимые вычисления дают:

1) оценки средних значений по каждой обучающей выборке (12.1), подсчитанные по формуле (12.11):

$$\hat{a}_1 = (576,0; 596,0)^T; \quad \hat{a}_2 = (598,46; 710,77)^T;$$

соответственно:

$$\hat{a}_1 - \hat{a}_2 = (-22,46; -114,77)^T; (\hat{a}_1 + \hat{a}_2)/2 = (587,23; 653,39)^T;$$

2) оценка общей ковариационной матрицы $\hat{\Sigma}$ производится по формуле (12.12) или, что то же:

$$\begin{aligned}\hat{\Sigma} &= \frac{1}{n_1 + n_2 - 2} \sum_{j=1}^2 \sum_{i=1}^{n_j} (X_{ji} - \hat{a}_j)(X_{ji} - \hat{a}_j)^T \\ &= \frac{1}{10 + 13 - 2} \left[\begin{pmatrix} 56240 & 17240 \\ 17240 & 33240 \end{pmatrix} + \begin{pmatrix} 121569 & 25615 \\ 25615 & 56492 \end{pmatrix} \right] \\ &= \begin{pmatrix} 8467 & 2041 \\ 2041 & 4273 \end{pmatrix};\end{aligned}$$

в результате обращения этой матрицы получаем

$$\hat{\Sigma}^{-1} = \begin{pmatrix} 0,0001335 & -0,0000637 \\ -0,0000637 & 0,0002645; \end{pmatrix}$$

$$3) \hat{\Sigma}^{-1}(\hat{a}_1 - \hat{a}_2) = \begin{pmatrix} 0,0043 \\ -0,0289 \end{pmatrix};$$

$$X_0 - \frac{1}{2}(\hat{a}_1 + \hat{a}_2) = \begin{pmatrix} 740 - 587,23 \\ 590 - 653,39 \end{pmatrix} = \begin{pmatrix} 152,77 \\ -63,39; \end{pmatrix}$$

4) поскольку численное значение выражения, определяемого соотношением (12.13'''), в нашем случае равно

$$\begin{pmatrix} 152,77 \\ -63,39 \end{pmatrix}^T \begin{pmatrix} 0,0043 \\ -0,0289 \end{pmatrix} = 2,489 > 0,$$

то наблюдение X_0 должно быть отнесено к классу 1, а это значит, что есть основания к тому, чтобы диагностировать анализируемую фирму как фирму, в той или иной форме уклоняющуюся от уплаты налогов.

12.3. Классификация без обучения (параметрический случай): расщепление смесей вероятностных распределений

В пп. 12.3 и 12.4 описаны методы классификации объектов (индивидуумов, семей, предприятий, городов, стран, технических систем, признаков

и т. д.) $O_1, O_2, \dots, O_n$ в ситуации, когда отсутствуют так называемые обучающие выборки (12.1), а исходная информация о классифицируемых объектах представлена либо в форме матрицы X «объект-свойство» (см. (9.1')), либо в форме матрицы парных сравнений объектов (см. (9.2), статический вариант), где величина $\gamma_{ij} = \rho_{ij}$ характеризует взаимную отдаленность (или близость) объектов O_i и O_j .

Переход от формы исходных данных типа «объект-свойство» к форме матрицы попарных расстояний осуществляется посредством задания способа вычисления расстояния (близости) между парой объектов, когда известны координаты (значения признаков) каждого из них.

В зависимости от наличия и характера априорных сведений о природе искомых классов и от конечных прикладных целей исследования следует обратиться либо к данному пункту 12.3, где описаны методы *расщепления смесей вероятностных распределений*, которые оказываются полезными в том случае, когда каждый (j -й) класс интерпретируется как параметрически заданная одномодальная генеральная совокупность $f_j(X; \Theta_j)$ ($j = 1, 2, \dots, k$) при неизвестном значении определяющего ее векторного значения параметра Θ_j и соответственно каждое из классифицируемых наблюдений X_i считается извлеченным из одной из этих (но не известно, из какой именно) генеральных совокупностей; либо п. 12.4, где описаны методы *автоматической классификации (кластер-анализа)* многомерных наблюдений, которыми исследователь вынужден пользоваться, когда не имеет оснований для интерпретации классифицируемых наблюдений в качестве выборки из какой-либо вероятностной генеральной совокупности или не располагает априорными сведениями, достаточными для параметрического представления искомых классов. В том же пункте кратко излагаются основные классификационные *процедуры иерархического типа*, используемые в ситуациях, когда «на выходе» исследователь хочет иметь не столько окончательный вариант разбиения анализируемой совокупности объектов на классы, сколько общее наглядное представление о стратификационной структуре этой совокупности (например, в виде специально устроенного графа — *дендрограммы*).

12.3.1. Понятие смеси вероятностных распределений

Начнем пояснение понятия смеси распределений с рассмотрения ряда конкретных примеров.

Пример 12.5. В отдел технического контроля (ОТК) поступают партии изделий, составленные с помощью случайного извлечения из объединенной продукции двух станков (станка A и станка B). Изделия кон-

тролируются по некоторому количественному параметру (линейному размеру) ξ мм, так что результатом контроля i -го изделия партии является число x_i мм. Изделия на станках не маркируются, так что в ОТК не известно, на каком именно станке произведено каждое из них. Производительность станка A в 1,5 раза выше производительность станка B . Задано номинальное значение контролируемого параметра $a = 65$ мм и известно, что точность работы станков характеризуется одинаковой величиной среднеквадратических отклонений $\sigma_A = \sqrt{D\xi_A}$ и $\sigma_B = \sqrt{D\xi_B}$, равной $1,0$ мм¹. Позже выяснилось, что станок A был настроен правильно (производил изделие со средним значением $E\xi_A = 65$ мм, равным номиналу), в то время как настройка станка B была сбита в направлении завышения номинала (а именно $E\xi_B = 67$ мм).

Известно также, что распределение размеров изделий, произведенных на каком-то определенном станке, описывается *нормальным законом* с параметрами $a_\gamma = E\xi_\gamma$ и $\sigma_\gamma^2 = D\xi_\gamma$ ($\gamma = A$ или $\gamma = B$).

Очевидно, анализируемая в ОТК по наблюдениям $x_1, x_2, \dots, x_n \dots$ генеральная совокупность будет состоять из *смеси двух нормальных генеральных совокупностей*, одна из которых представляет продукцию станка A и описывается в соответствии с вышесказанным плотностью

$$f_A(x) = \varphi(x; a_A, \sigma_A^2) = \frac{1}{\sqrt{2\pi\sigma_A^2}} e^{-\frac{(x-a_A)^2}{2\sigma_A^2}},$$

а другая — продукцию станка B и описывается плотностью

$$f_B(x) = \varphi(x; a_B, \sigma_B^2) = \frac{1}{\sqrt{2\pi\sigma_B^2}} e^{-\frac{(x-a_B)^2}{2\sigma_B^2}}.$$

Обозначая $\Theta_\gamma = (a_\gamma, \sigma_\gamma^2)$, а удельный вес изделий станка γ через π_γ ($\gamma = A, B$), можем записать уравнение функции плотности $f(x)$, описывающей закон распределения анализируемого признака ξ во всей (*объединенной*) генеральной совокупности, в виде:

$$f(x) = \pi_A \varphi(x; \Theta_A) + \pi_B \varphi(x; \Theta_B). \quad (12.14)$$

Учитывая, что в объединенной генеральной совокупности продукции станка A в 1,5 раза больше, чем продукции станка B (поскольку производительность станка A в 1,5 раза выше), а также то, что $a_A = 65$ мм,

¹ Случайная величина ξ и ее числовые характеристики будут снабжаться нижним индексом (A или B) в тех случаях, когда речь идет о продукции какого-то определенного станка (соответственно станка A или станка B).

$a_B = 67$ мм, $\sigma_A^2 = \sigma_B^2 = 1$ мм², имеем:

$$f(x) = 0,6 \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-65)^2}{2}} + 0,4 \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-67)^2}{2}}. \quad (12.14')$$

Правыми частями уравнений (12.14) и (12.14') и представлен частный случай того, что принято называть *смесью вероятностных распределений*¹.

На рис. 12.1 представлены графики функций плотности $f_A(x)$, $f_B(x)$ и $f(x)$.

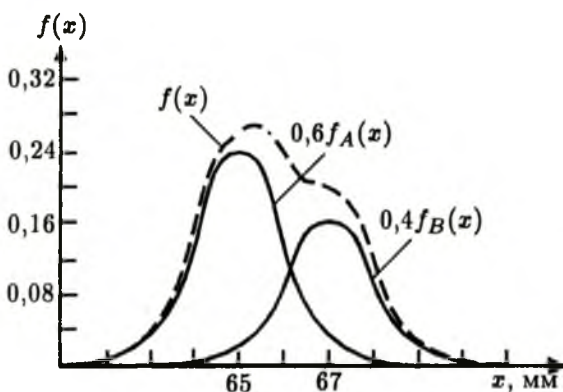


Рис. 12.1. Графики функций плотности отдельных компонентов и самой смеси из примера 12.5

В соотношениях (12.14) и (12.14') величины $\pi_A = 0,6$ и $\pi_B = 0,4$ представляют удельные веса соответствующих компонентов смеси (их еще называют *априорными вероятностями* появления наблюдений из данного компонента смеси), а $\Theta_A = (a_A, \sigma_A^2)$ и $\Theta_B = (a_B, \sigma_B^2)$ — векторные параметры, от значений которых зависят законы распределения компонентов смеси.

Если сотрудники ОТК или потребители изделий-полуфабрикатов захотят по наблюдениям $x_1, x_2, \dots$ определить, *на каком именно станке произведено каждое из них*, то как раз и возникает одна из типичных задач классификации наблюдений в условиях отсутствия обучающих выборок (конечно, в данном примере можно представить себе специально

¹ Речь идет о частном случае, поскольку в общей модели смеси распределений, во-первых, могут участвовать более чем два (и даже континуум) составляющих смесь распределений, а во-вторых, анализируемые распределения могут быть многомерными и не обязаны быть однотипными (в данном примере оба компонента нормальные).

организованное производство этих изделий, в результате которого можно получить отдельно изделия от станка A и отдельно — от станка B и использовать их в дальнейшем в качестве обучающих выборок).

Пример 12.6. Речь идет о выявлении и анализе типов потребительского поведения по данным обследований семейных бюджетов (см. выше, пример 12.1). Здесь представлен один из фрагментов исследования, проведенного с целью изучения (на базе семейных бюджетов) дифференциации потребностей, выявления основных типов потребительского поведения и определения главных типобразующих признаков (социально-демографической, региональной, экономической природы). Исследуемым многомерным признаком является вектор Y показателей $y^{(1)}, y^{(2)}, \dots, y^{(p)}$ потребительского поведения семьи, т. е. каждой (i -й) обследованной семье ставится в соответствие многомерное наблюдение

$$Y_i = \begin{pmatrix} y_i^{(1)} \\ \vdots \\ y_i^{(p)} \end{pmatrix}, \quad i = 1, 2, \dots, n,$$

где $y_i^{(m)}$ — удельное (т. е. рассчитанное в среднем на одного члена семьи) количество m -го вида благ (товаров или услуг, включая сбережения), потребляемое i -й обследованной семьей за «единицу времени» (например, за год) и выраженное в натуральных или денежных единицах.

В соответствии с одним из принятых в исследовании базовых исходных допущений постулируется существование в анализируемом пространстве $\Pi^p(Y)$ ($Y \in \Pi^p(Y)$) сравнительно небольшого (и неизвестного) числа k типов потребительского поведения, таких, что различия в структуре потребления Y семей одного типа носят *случайный характер* (т. е. обусловлены влиянием множества случайных, не поддающихся управлению и учету факторов) и *незначительны* по сравнению с различиями в потребительском поведении семей, представляющих разные типы. При этом предполагается, что случайный разброс структур потребительских поведений $Y(j)$ *внутри* любого (j -го) типа описывается многомерным (в нашем случае p -мерным) *нормальным* законом распределения с некоторым вектором средних (и в то же время — наиболее характерных, наиболее часто наблюдаемых) значений

$$a(j) = \begin{pmatrix} a^{(1)}(j) \\ a^{(2)}(j) \\ \vdots \\ a^{(p)}(j) \end{pmatrix}$$

и с ковариационной матрицей

$$\Sigma(j) = \begin{pmatrix} \sigma_{11}(j) & \sigma_{12}(j) & \dots & \sigma_{1p}(j) \\ \sigma_{21}(j) & \sigma_{22}(j) & \dots & \sigma_{2p}(j) \\ \dots & \dots & \dots & \dots \\ \sigma_{p1}(j) & \sigma_{p2}(j) & \dots & \sigma_{pp}(j) \end{pmatrix}$$

(см. сведения о многомерном нормальном законе в п. 3.1.5, формула (3.10)).

Однако в начале исследования нет сведений об упомянутых гипотетических типах потребительского поведения: неизвестно ни их число k , ни значения определяющих эти типы многомерных параметров $\Theta_j = (\mathbf{a}(j), \Sigma(j))$. Поэтому мы вынуждены рассматривать имеющиеся в нашем распоряжении результаты бюджетных обследований семей

$$Y_1, Y_2, \dots, Y_n \quad (12.15)$$

как выборку из генеральной совокупности, являющейся смесью многомерных нормальных законов распределения. Другими словами, функция плотности $f(Y)$, описывающая распределение вектора Y в этой объединенной генеральной совокупности, имеет вид

$$f(Y) = \sum_{j=1}^k \pi_j \varphi(Y; \mathbf{a}(j); \Sigma(j)), \quad (12.16)$$

где π_j ($j = 1, 2, \dots, k$) — не известный нам удельный вес (априорная вероятность) семей j -го типа потребительского поведения в общей совокупности семей;

$$\varphi(Y; \mathbf{a}(j); \Sigma(j)) = \frac{1}{(2\pi)^{\frac{k}{2}} |\Sigma(j)|^{\frac{1}{2}}} e^{-\frac{1}{2}(Y-\mathbf{a}(j))^T \Sigma(j)^{-1} (Y-\mathbf{a}(j))} -$$

многомерная нормальная плотность, описывающая закон распределения исследуемого признака $Y(j)$ внутри совокупности семей j -го типа потребительского поведения ($j = 1, 2, \dots, k$).

Далее необходимо по выборке (12.15) оценить неизвестные значения параметров k , $\pi_1, \pi_2, \dots, \pi_{k-1}$, $\mathbf{a}(j)$ и $\Sigma(j)$ ($j = 1, \dots, k$) модели (12.16), чтобы в конечном счете суметь расклассифицировать (в определенном смысле наилучшим образом) семьи 12.15 по искомым типам потребительского поведения. Общая схема действий, увязывающая задачу статистического оценивания параметров смеси типа (12.16) с задачей автоматической классификации, изложена в п. 12.3.2.

Общая математическая модель смеси распределений. Рассмотренные в примерах смеси (12.14) и (12.16) представляют собой частные случаи общей модели смеси. Обобщение рассмотренных в примерах смесей может быть произведено в направлении: 1) отказа от конечности и даже дискретности компонентов, составляющих смесь, распространения понятия смеси на непрерывную смешивающую функцию; 2) отказа однотипности участвующих в смеси компонентов (под однотипностью компонентов-распределений понимается их принадлежность к общему параметрическому семейству распределений, например к нормальному).

Итак, пусть имеется двухпараметрическое семейство p -мерных плотностей (полигонов вероятностей) распределения

$$F = \{f_w(X; \Theta(\omega))\}, \quad (12.17)$$

где одномерный (целочисленный или непрерывный) параметр ω в качестве нижнего индекса функции f определяет *специфику общего вида* каждого компонента-распределения в смеси, а в качестве аргумента при многомерном, вообще говоря, параметре Θ определяет зависимость значений хотя бы части компонентов этого параметра от того, в каком именно составляющем распределении f_ω он присутствует. И пусть

$$\Psi = \{\psi(\omega)\} \quad (12.18)$$

— семейство *смешивающих функций распределения*.

Будем, как и ранее, понимать под $f(X)$ *функцию плотности*, если исследуется *непрерывная* случайная величина, и *полигон вероятностей* $P\{\xi = X\}$, если анализируемая случайная величина дискретна. Аналогично значение функции $\psi(\omega) = \pi_\omega$ ($\omega = 1, 2, \dots, k$) интерпретируется как априорная вероятность появления элементов класса ω , если смесь состоит из *конечного* числа (k) компонентов $f_\omega(X; \Theta(\omega))$, и $\psi(\omega)$ — функция плотности, если смесь состоит из континуального множества компонентов.

Функция $f(X)$ называется смесью вероятностных распределений (дискретной или непрерывной), если она представима в виде соответственно

$$f(X) = \sum_{j=1}^k \pi_j f_j(X; \Theta(j)) \quad (12.19)$$

или

$$f(X) = \int f_\omega(X; \Theta(\omega)) \psi(\omega) d\omega. \quad (12.19')$$

В данном пункте нас интересует использование моделей смесей в теории и практике автоматической классификации, поэтому сузим данное выше определение смеси и будем рассматривать в дальнейшем

лишь *случай конечного числа* k возможных значений параметра ω , что соответствует ситуации, в которой величины $\psi(\omega)$ будут играть роль удельных весов (априорных вероятностей) π_j компонентов смеси ($j = 1, 2, \dots, k$)¹. Если же дополнительно постулировать *однотипность* компонентов-распределений $f_j(X; \Theta(j))$, т. е. принадлежность всех $f_j(X; \Theta(j))$ к одному общему семейству $\{\tilde{f}(X; \Theta)\}$, то модель смеси может быть представлена в виде

$$f(X) = \sum_{j=1}^k \pi_j \tilde{f}(X; \Theta(j)). \quad (12.19'')$$

Интерпретация в задачах автоматической классификации j -го компонента смеси (j -й генеральной совокупности) в качестве j -го искомого класса (сгустка, скопления) обуславливает естественность дополнительного условия, накладываемого на плотности (полигоны вероятностей) $f_j(X; \Theta(j))$ и заключающегося в *их одномодальности*.

12.3.2. Задача расщепления смеси распределений

Решить задачу расщепления смеси распределений — это значит по выборке классифицируемых наблюдений

$$X_1, X_2, \dots, X_n, \quad (12.20)$$

извлеченной из генеральной совокупности, являющейся смесью (12.19'') генеральных совокупностей типа (12.17) (при заданном общем виде составляющих смесь функций $f_\omega(X; \Theta(\omega))$), построить статистические оценки для числа компонентов смеси k , их удельных весов (априорных вероятностей) $\pi_1, \pi_2, \dots, \pi_k$ и, главное, для каждого из компонентов $f_\omega(X; \Theta(\omega))$ анализируемой смеси (12.19''). В некоторых частных случаях имеющиеся

¹ Модель смеси вида (12.19') с непрерывной смешивающей функцией $\psi(\omega)$ также является полезным инструментом социально-экономического анализа и моделирования. Так, например, при моделировании распределения населения по величине среднедушевого дохода каждый компонент смеси интерпретируется как население *однородной* (по источникам формирования доходов, территориальным и социально-профессиональным признакам) страты, а все население — как *континуальная смесь* таких страт. Подобный подход лежит в основе обоснования логарифмически-нормальной модели распределения населения по доходу (см. Айвазян С. А. Модель формирования распределения населения России по величине среднедушевого дохода // Экономика и математические методы, том 33 (1997), № 4, с. 74–86).

априорные сведения дают исследователю точное знание числа компонентов смеси k , а иногда и априорных вероятностей $\pi_1, \pi_2, \dots, \pi_k$. Тогда задача расщепления смеси сводится лишь к оцениванию функций $f_\omega(X; \Theta(\omega))$.

Отметим, что задача расщепления сохраняет смысл и *применительно к генеральным совокупностям*, т.е. в теоретическом варианте. В этом случае она заключается в восстановлении компонентов $f_\omega(X; \theta(\omega))$ и смешивающей функции $\psi(\omega)$ по заданной левой части $f(X)$ в соотношениях (12.19) или (12.19') и называется *задачей идентификации компонентов смеси*.

12.3.3. Общая схема решения задачи автоматической классификации в рамках модели смеси распределений (сведение к схеме дискриминантного анализа)

Базовая идея, лежащая в основе принятия решения, к какой из k анализируемых генеральных совокупностей отнести данное классифицируемое наблюдение X_i , состоит в том, что *наблюдение следует отнести к той генеральной совокупности, в рамках которой оно выглядит наиболее правдоподобным*. Другими словами, если дано точное описание (например, в виде функций $f_1(X), \dots, f_k(X)$ плотности в непрерывном случае или полигонов вероятностей в дискретном) конкурирующих генеральных совокупностей, то следует поочередно вычислить значения функций правдоподобия для данного наблюдения X_i в рамках каждой из рассматриваемых генеральных совокупностей (т.е. вычислить значения $f_1(X_i), f_2(X_i), \dots, f_k(X_i)$) и отнести X_i к тому классу, функция правдоподобия которого максимальна¹. Если же известен лишь общий вид функций $f_1(X; \Theta_1), f_2(X; \Theta_2), \dots, f_k(X; \Theta_k)$, описывающих анализируемые классы, но не известны значения, вообще говоря, многомерных параметров $\Theta_1, \Theta_2, \dots, \Theta_k$, и если при этом располагают так называемыми обучающими выборками, то данный случай лежит в рамках *параметрической схемы дискриминантного анализа* (ДА) и порядок действий будет следующим (п. 12.2): сначала по j -й обучающей выборке оцениваем параметр Θ_j ($j = 1, 2, \dots, k$), а затем производим классификацию наблюдений, руководствуясь тем же самым принципом максимального правдоподобия, что и в случае полностью известных функций $f_j(X)$.

В схеме *автоматической классификации*, опирающейся на модель

¹ Для большей ясности здесь подразумевается простой случай равных априорных вероятностей и равных потерь от неправильного отнесения наблюдения X_i к любому из классов. Более общая схема и более подробно представлена в п. 12.2.

смеси распределений, как и в схеме параметрического ДА, задающие искомые классы функции $f_1(X; \Theta_1)$, $f_2(X; \Theta_2), \dots$ также известны лишь с точностью до значений параметров. Но в схеме автоматической классификации неизвестные значения параметров $\Theta_1, \Theta_2, \dots$, так же, впрочем, как и параметров $k, \pi_1, \pi_2, \dots, \pi_k$, *оцениваются не по обучающим выборкам (их нет в распоряжении исследователя), а по классифицируемым наблюдениям $X_1, X_2, \dots, X_n$ с помощью одного из известных методов статистического оценивания параметров (метода максимального правдоподобия, метода моментов или какого-либо другого). Начиная с момента, когда по выборке (12.20) сумели получить оценки $\hat{k}, \hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_k, \hat{\Theta}_1, \dots, \hat{\Theta}_k$ неизвестных параметров $k, \pi_1, \pi_2, \dots, \pi_k, \Theta_1, \Theta_2, \dots, \Theta_k$ модели (12.19''), снова имеем схему дискриминантного анализа и собственно процесс классификации наблюдений (12.20) производим точно так же, как и в схеме параметрического ДА (т.е. относим наблюдение X_i к классу с номером j_0 , если $\hat{\pi}_{j_0} f_{j_0}(X_i; \hat{\Theta}_{j_0}) = \max_{1 \leq j \leq k} \{\hat{\pi}_j f_j(X_i; \hat{\Theta}_j)\}$.*

Итак, *главное отличие схемы параметрического ДА от схемы автоматической классификации, производимой в рамках модели смеси распределений, — в способе оценивания неизвестных параметров, от которых зависят функции, описывающие классы. Но оценивание параметров в модели смеси — процесс неизмеримо более сложный, чем оценивание параметров по обучающим выборкам. Весьма подробное описание процедур оценивания параметров смеси вероятностных распределений читатель найдет в книге [Айвазян С. А. и др., 1989].*

12.4. Классификация без обучения (непараметрический случай): методы кластер-анализа

12.4.1. Общая постановка задачи автоматической классификации

Итак, как и в п. 12.3, мы *не располагаем обучающими выборками* и имеем лишь n подлежащих классификации наблюдений, заданных либо матрицей X (9.1'), либо матрицей γ (9.2) (в *статичном* варианте), содержащей все попарные расстояния (меры близости) между классифицируемыми наблюдениями. Однако *в отличие от п. 12.3 в данном случае отсутствует и априорная информация о характере распределения наблюдений X_i внутри каждого из классов* (в предыдущем пункте нам был

известен общий параметрический вид этих распределений). Теперь же мы можем располагать лишь самыми общими сведениями об этих распределениях. Эти сведения могут относиться, например, к компактности или ограниченности диапазонов изменения компонент классифицируемых многомерных наблюдений, к свойствам непрерывности (гладкости) соответствующих законов распределения вероятностей. Заметим, что параллельно можно обсуждать задачу классификации *признаков* $X^{(j)}$ (а не объектов X_i), с той лишь разницей, что каждый из признаков задается соответствующей *строкой* матрицы X . В дальнейшем, если это специально не оговорено, не будем разделять изложение этой проблемы на «объекты» и «признаки», поскольку все постановки задач и основная методологическая схема исследования здесь общие.

В общей (нестрогой) постановке *проблема автоматической классификации объектов заключается в том, чтобы всю анализируемую совокупность объектов* $O = \{O_i\}$ ($i = \overline{1, n}$), *статистически представленную в виде матриц* X *или* γ , *разбить на сравнительно небольшое число (заранее известное или нет) однородных, в определенном смысле, групп или классов.*

Для формализации этой проблемы удобно интерпретировать анализируемые объекты в качестве точек в соответствующем признаковом пространстве. Если исходные данные представлены в форме матрицы (X), то эти точки являются непосредственным геометрическим изображением многомерных наблюдений $X_1, X_2, \dots, X_n$ в p -мерном пространстве $\Pi^p(X)$ с координатными осями $Ox^{(1)}, Ox^{(2)}, \dots, Ox^{(p)}$. Если же исходные данные представлены в форме матрицы попарных взаимных расстояний γ , то исследователю не известны непосредственно координаты этих точек, но зато задана структура попарных расстояний (близостей) между объектами. Естественно предположить, что геометрическая близость двух или нескольких точек в этом пространстве означает близость «физических» состояний соответствующих объектов, их однородность. Тогда проблема классификации состоит в разбиении анализируемой совокупности точек — наблюдений на сравнительно небольшое число (заранее известное или нет) классов таким образом, чтобы объекты, принадлежащие одному классу, находились бы на сравнительно небольших расстояниях друг от друга. Полученные в результате разбиения классы часто называют *кластерами* (таксонами, образами)¹, а методы их нахождения соответственно кластер-

¹ Cluster (англ.) — гроздь, пучок, скопление, группа элементов, характеризуемых каким-либо общим свойством. Тахон (англ.) — систематизированная группа любой категории (термин биологического происхождения). Название «кластер-анализ»

анализом, численной таксономией распознаванием образов с самообучением.

Однако, берясь за решение задачи классификации, исследователь с самого начала должен четко представлять, какую именно из двух задач он решает. Рассматривает ли он обычную задачу разбиения статистически обследованного (p -мерного) диапазона изменения значений анализируемых признаков на *интервалы (гиперобласти) группирования*, в результате решения которой исследуемая совокупность объектов разбивается на некоторое число групп так, что объекты такой одной группы находятся друг от друга на сравнительно небольшом расстоянии (многомерный аналог задачи построения интервала группирования при обработке одномерных наблюдений см. «Переход к группированным данным» в п. 6.1). Либо он пытается определить *естественное расслоение* исходных наблюдений на четко выраженные кластеры, сгустки, лежащие друг от друга на некотором расстоянии, но не разбивающиеся на столь же удаленные части. В вероятностной интерпретации (т. е. если интерпретировать классифицируемые наблюдения $X_1, X_2, \dots, X_n$ как выборку из некоторой многомерной генеральной совокупности, описываемой функцией плотности или полигоном распределения $f(X)$, как правило, не известными исследователю) вторая задача может быть сформулирована как задача выявления областей повышенной плотности наблюдений, т. е. таких областей возможных значений анализируемого многомерного признака X , которые *соответствуют локальным максимумам функции $f(X)$* .

Если первая задача — построения областей группирования — всегда имеет решение, то при второй постановке результат может быть и отрицательным: может оказаться, что множество исходных наблюдений не обнаруживает естественного расслоения на кластеры (например, образует один общий кластер).

Из методологических соображений (в частности, для упрощения понимания читателем некоторых основных идей теории автоматической классификации и для создания удобной схемы исследования свойств различных классификационных процедур) будем иногда вводить в рассмотрение теоретические вероятностные характеристики анализируемой совокупности: генеральную совокупность, плотность (полигон) распределения, теоретические средние значения, дисперсии, ковариации и т. п. Очевидно, если мысленно «продолжить» множество классифицируемых наблюдений *до всей генеральной совокупности*, задача классификации заключа-

для совокупности методов решения задач такого типа было впервые введено, по-видимому, Трайном в 1939 г. (Tryon R. C. Cluster Analysis//Ann. Arb., Edw. Brothers. 1939).

ется в разбиении анализируемого признакового пространства $\Pi^P(X)$ на некоторое число непересекающихся областей. Условимся в дальнейшем называть такую схему *теоретико-вероятностной модификацией задачи кластер-анализа*.

12.4.2. Расстояния между отдельными объектами и меры близости объектов друг к другу

Наиболее труден и наименее формализован в задаче автоматической классификации момент, связанный с определением *понятия однородности объектов*.

В общем случае понятие однородности объектов определяется заданием правила вычисления величины ρ_{ij} , характеризующей либо расстояние $d(O_i, O_j)$ между объектами O_i и O_j из исследуемой совокупности $O(i, j = 1, 2, \dots, n)$, либо степень близости (сходства) $r(O_i, O_j)$ тех же объектов. Если задана функция $d(O_i, O_j)$, то близкие в смысле этой метрики объекты считаются однородными, принадлежащими к одному классу. Естественно, при этом необходимо сопоставление $d(O_i, O_j)$ с некоторым пороговым значением, определяемым в каждом конкретном случае по-своему.

При задании расстояний и мер близости нужно помнить о необходимости соблюдения следующих естественных требований: требования *симметрии* ($d(O_i, O_j) = d(O_j, O_i)$ и $r(O_i, O_j) = r(O_j, O_i)$); требования *максимального сходства объекта с самим собой* ($r(O_i, O_i) = \max_{1 \leq j \leq n} r(O_i, O_j)$) и требования при заданной метрике *монотонного убывания* $r(O_i, O_j)$ по $d(O_i, O_j)$, т. е. из $d(O_k, O_l) \geq d(O_i, O_j)$ должно с необходимостью следовать выполнение неравенства $r(O_k, O_l) \leq r(O_i, O_j)$.

Конечно, *выбор метрики (или меры близости) является узловым моментом исследования*, от которого решающим образом зависит окончательный вариант разбиения объектов на классы при заданном алгоритме разбиения. В каждой конкретной задаче этот выбор должен производиться по-своему. При этом решение данного вопроса зависит в основном от главных целей исследования, физической и статистической природы вектора наблюдений X , полноты априорных сведений о характере вероятностного распределения X . Так, например, если из конечных целей исследования и из природы вектора X следует, что понятие однородной группы естественно интерпретировать как генеральную совокупность с одновершинной плотностью (полигоном частот) распределения, и если к тому же известен общий вид этой плотности, то следует воспользоваться общим подходом, описанным в п. 12.3. Если, кроме того, известно, что наблюде-

ния X_i извлекаются из *нормальных* генеральных совокупностей с *одной и той же матрицей ковариаций*, то естественной мерой отдаленности двух объектов друг от друга является *расстояние маталанобисского типа* (см. ниже).

В качестве примеров расстояний и мер близости, сравнительно широко используемых в задачах кластер-анализа, приведем здесь следующие.

Общий вид метрики маталанобисского типа. В общем случае зависимых компонент $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ вектора наблюдений X и их различной значимости в решении вопроса об отнесении объекта (наблюдения) к тому или иному классу обычно пользуются *обобщенным* («взвешенным») *расстоянием маталанобисского типа*, задаваемым формулой¹

$$d_0(X_i, X_j) = \sqrt{(X_i - X_j)^T \Delta^T \Sigma^{-1} \Delta (X_i - X_j)}.$$

Здесь Σ — ковариационная матрица генеральной совокупности, из которой извлекаются наблюдения X_i , а Δ — некоторая симметричная неотрицательно-определенная матрица «весовых» коэффициентов λ_{mq} , которая чаще всего выбирается диагональной.

Следующие три вида расстояний, хотя и являются частными случаями метрики d_0 , все же заслуживают специального описания.

Обычное евклидово расстояние

$$d_E(X_i, X_j) = \sqrt{\sum_{k=1}^p (x_i^{(k)} - x_j^{(k)})^2}.$$

К ситуациям, в которых использование этого расстояния можно признать оправданным, прежде всего относят следующие:

- наблюдения X извлекаются из генеральных совокупностей, описываемых многомерным нормальным законом с ковариационной матрицей вида $\sigma^2 \cdot I$, т. е. компоненты X взаимно независимы и имеют одну и ту же дисперсию;

- компоненты $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ вектора наблюдений X однородны по своему физическому смыслу, причем установлено, например с помощью опроса экспертов, что все они одинаково важны с точки зрения решения вопроса об отнесении объекта к тому или иному классу;

¹ В случаях, когда каждый объект O_i представлен вектором признаков X_i (т. е. в случае исходных данных, представленных в форме X), часто удобнее в формулах и различных соотношениях вместо O_i писать сразу X_i . Например, $d(X_i, X_j)$ вместо $d(O_i, O_j)$.

• признаковое пространство совпадает с геометрическим пространством нашего бытия, что может быть лишь в случаях $p = 1, 2, 3$, и понятие близости объектов соответственно совпадает с понятием геометрической близости в этом пространстве, например классификация попаданий при стрельбе по цели.

Взвешенное евклидово расстояние

$$d_{\text{вЕ}}(X_i, X_j) = \sqrt{\sum_{k=1}^p \omega_k (x_i^{(k)} - x_j^{(k)})^2}.$$

Обычно применяется в ситуациях, в которых так или иначе удастся приписать каждой из компонент $x^{(k)}$ вектора наблюдений X некоторый неотрицательный «вес» ω_k , пропорциональный степени его важности с точки зрения решения вопроса об отнесении заданного объекта к тому или иному классу. Удобно полагать при этом $0 \leq \omega_k \leq 1$, $k = \overline{1, p}$.

Определение весов ω_k связано, как правило, с дополнительным исследованием, например получением и использованием обучающих выборок, организацией опроса экспертов и обработкой их мнений, использованием некоторых специальных моделей. Попытки определения весов ω_k только по информации, содержащейся в исходных данных, как правило, не дают желаемого эффекта, а иногда могут лишь отдалить от истинного решения. Достаточно заметить, что в зависимости от весьма тонких и незначительных вариаций физической и статистической природы исходных данных можно привести одинаково убедительные доводы в пользу двух диаметрально противоположных решений этого вопроса: выбирать ω_k пропорционально величине среднеквадратической ошибки признака $x^{(k)}$ либо пропорционально обратной величине среднеквадратической ошибки этого же признака.

Хеммингово расстояние. Используется как мера различия объектов, задаваемых *дихотомическими* признаками. Оно задается с помощью формулы

$$d_H(X_i, X_j) = \sum_{s=1}^p |x_i^{(s)} - x_j^{(s)}|$$

и, следовательно, равно числу v_{ij} несовпадений значений соответствующих признаков в рассматриваемых i -м и j -м объектах.

Другие меры близости для дихотомических признаков. Меры близости объектов, описываемых набором дихотомических признаков, обычно основаны на характеристиках $v_{ij}^{(0)}$, $v_{ij}^{(1)}$ и $v_{ij} = v_{ij}^{(0)} + v_{ij}^{(1)}$, где

$v_{ij}^{(0)}(v_{ij}^{(1)})$ — число нулевых (единичных) компонент, совпавших в объектах X_i и X_j . Так, например, если из каких-либо профессиональных соображений или априорных сведений следует, что все p признаков исследуемых объектов можно считать равноправными, а эффект от совпадения или несовпадения нулей такой же, как и от совпадения или несовпадения единиц, то в качестве меры близости объектов X_i и X_j используют величину $r(X_i, X_j) = \frac{v_{ij}}{p}$.

Весьма полный обзор различных мер близости объектов читатель найдет, например, в статье Г.В. Раушенбаха «Проблемы измерения близости в задачах анализа данных» (сборник «Программно-алгоритмическое обеспечение анализа данных в медико-биологических исследованиях». — М.: Наука, 1987, с. 41–54).

О физически содержательных мерах близости объектов. В некоторых задачах классификации объектов, не обязательно описываемых количественно, естественнее использовать в качестве меры близости объектов (или расстояния между ними) некоторые физически содержательные числовые параметры, так или иначе характеризующие взаимоотношения между объектами. Примером может служить задача классификации с целью агрегирования отраслей экономики, решаемая на основе матрицы межотраслевого баланса. Классифицируемым объектом в данном примере является отрасль экономики, а матрица межотраслевого баланса представлена элементами s_{ij} , где под s_{ij} подразумевается сумма годовых поставок в денежном выражении i -й отрасли в j -ю. В качестве матрицы близости (r_{ij}) в этом случае естественно взять, например, *симметризованную нормированную* матрицу межотраслевого баланса. При этом под *нормировкой* понимается преобразование, при котором денежное выражение поставок из i -й отрасли в j -ю заменяется долей этих поставок по отношению ко всем поставкам i -й отрасли. Симметризацию же нормированной матрицы межотраслевого баланса можно проводить различными способами. Так, например, близость между i -й и j -й отраслями можно выразить через *среднее* значение их взаимных нормированных поставок.

О мерах близости числовых признаков (отдельных факторов). Решение задач классификации многомерных данных, как правило, предусматривает в качестве предварительного этапа исследования реализацию методов, позволяющих существенно сократить размерность исходного факторного пространства, выбрать из компонент $x^{(1)}, \dots, x^{(p)}$ наблюдаемых векторов X сравнительно небольшое число наиболее существенных, наиболее информативных. Для этих целей бывает полезно рассмотреть каждую из компонент $x^{(1)}, \dots, x^{(p)}$ в качестве объекта, подлежа-

щего классификации. Дело в том, что разбиение признаков $x^{(1)}, \dots, x^{(p)}$ на небольшое число однородных в некотором смысле групп позволит исследователю сделать вывод, что компоненты, входящие в одну группу, в определенном смысле сильно связаны друг с другом и несут информацию о *каком-то одном свойстве* исследуемого объекта. Следовательно, можно надеяться, что не будет большого ущерба в информации, если для дальнейшего исследования оставим лишь по одному представителю от каждой такой группы (процедура, реализующая эту идею, описана в п. 13.4.2).

Чаще всего в подобных ситуациях в качестве мер близости между отдельными признаками $x^{(i)}$ и $x^{(j)}$, так же как и между наборами таких признаков, используются различные характеристики степени их коррелированности и в первую очередь коэффициенты корреляции. Проблеме сокращения размерности анализируемого признакового пространства специально посвящена гл. 13 данного учебника.

12.4.3. Расстояния между классами объектов

При конструировании различных процедур классификации (кластер-процедур) в ряде ситуаций оказывается целесообразным введение понятия *расстояния между целыми группами объектов*. Приведем примеры наиболее распространенных расстояний, характеризующих взаимное расположение отдельных групп объектов. Пусть S_i — i -я группа (класс, кластер) объектов, n_i — число объектов, образующих группу S_i , вектор $\bar{X}(i)$ — среднее арифметическое векторных наблюдений, входящих в S_i (другими словами, $\bar{X}(i)$ — «центр тяжести» i -й группы), а $\rho(S_l, S_m)$ — расстояние между группами S_l и S_m .

Ниже приводятся наиболее употребительные и наиболее общие расстояния между классами объектов.

Расстояние, измеряемое по принципу «ближнего соседа» («nearest neighbour»)

$$\rho_{\min}(S_l, S_m) = \min_{X_i \in S_l, X_j \in S_m} d(X_i, X_j).$$

Расстояние, измеряемое по принципу «дальнего соседа» («furthest neighbour»)

$$\rho_{\max}(S_l, S_m) = \max_{X_i \in S_l, X_j \in S_m} d(X_i, X_j).$$

Расстояние, измеряемое по «центрам тяжести» групп

$$\rho(S_l, S_m) = d(\bar{X}(l), \bar{X}(m)).$$

Расстояние, измеряемое по принципу «средней связи» определяется как арифметическое среднее всевозможных попарных расстояний между представителями рассматриваемых групп, т. е.

$$\rho_{\text{cp}}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{X_j \in S_m} d(X_i, X_j).$$

Естественно задать вопрос: нельзя ли получить достаточно общую формулу, определяющую расстояние между классами по заданному расстоянию между отдельными элементами (наблюдениями), которая включила бы в себя в качестве частных случаев все рассмотренные выше виды расстояний?

Изыщное обобщение такого рода, основанное на понятии так называемого «обобщенного среднего», а точнее, степенного среднего, было предложено А. Н. Колмогоровым. Под обобщенным средним величин $c_1, c_2, \dots, c_N$ понимается выражение вида $M^F(c_1, c_2, \dots, c_N) = F^{-1}\left(\frac{1}{N} \sum_{i=1}^N F(c_i)\right)$, в котором $F(u)$ — некоторая функция и соответственно F^{-1} — преобразование, обратное к F . Частным случаем обобщенного среднего является *степенное среднее*, определяемое как

$$M_\tau(c_1, c_2, \dots, c_N) = \left(\frac{1}{N} \sum_{i=1}^N c_i^\tau \right)^{\frac{1}{\tau}}.$$

Нетрудно показать, что (при $c_i > 0, i = 1, 2, \dots, N$)

$$M_{-\infty}(c_1, c_2, \dots, c_N) = \min_{1 \leq i \leq N};$$

$$M_{+\infty}(c_1, c_2, \dots, c_N) = \max_{1 \leq i \leq N};$$

$$M_0(c_1, c_2, \dots, c_N) = \left(\prod_{i=1}^N c_i \right)^{1/N} \text{ — геометрическое среднее;}$$

$$M_1(c_1, c_2, \dots, c_N) = \frac{1}{N} \sum_{i=1}^N c_i \text{ — арифметическое среднее.}$$

Обобщенное (по Колмогорову) расстояние между классами, или обобщенное K -расстояние, вычисляется по формуле

$$\rho_\tau^{(K)}(S_l, S_m) = \left[\frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{X_j \in S_m} d^\tau(X_i, X_j) \right]^{\frac{1}{\tau}}. \quad (12.21)$$

В частности, при $\tau \rightarrow \infty$ при $\tau \rightarrow -\infty$ имеем

$$\rho_\infty^{(K)}(S_l, S_m) = \rho_{\max}(S_l, S_m); \quad \rho_{-\infty}^{(K)}(S_l, S_m) = \rho_{\min}(S_l, S_m).$$

Очевидно также

$$\rho_1^{(K)}(S_l, S_m) = \rho_{cp}(S_l, S_m).$$

Из (12.21) следует, что если $S(m, q) = S_m \cup S_q$ — группа элементов, полученная путем объединения кластеров S_m и S_q , то обобщенное K -расстояние между кластерами S_l и $S(m, q)$ определяется формулой

$$\rho_\tau^{(K)}(S_l, S(m, q)) = \left\{ \frac{n_m [\rho_\tau^{(K)}(S_l, S_m)]^\tau + n_q [\rho_\tau^{(K)}(S_l, S_q)]^\tau}{n_m + n_q} \right\}^{\frac{1}{\tau}}.$$

Понятие расстояния между группами элементов особенно важно в так называемых *агломеративных иерархических кластер-процедурах* (см. ниже п. 12.4.7), поскольку принцип работы таких алгоритмов состоит в последовательном объединении элементов, а затем и целых групп, сначала самых близких, а потом все более и более отдаленных друг от друга. Учитывая специфику подобных процедур, для задания расстояния между классами оказывается достаточным определить порядок пересчета расстояния между классом S_l и классом $S(m, q) = S_m \cup S_q$, являющимся объединением двух других классов S_m и S_q , по расстояниям $\rho_{lm} = \rho(S_l, S_m)$, $\rho_{lq} = \rho(S_l, S_q)$ и $\rho_{mq} = \rho(S_m, S_q)$ между этими классами. С этой целью часто используется также следующая общая формула¹ для вычисления расстояния между некоторым классом S_l и классом (m, q) :

$$\rho_{l(m, q)} = \rho(S_l, S(m, q)) = \alpha \rho_{lm} + \beta \rho_{lq} + \gamma \rho_{mq} + \delta |\rho_{lm} - \rho_{lq}|, \quad (12.22)$$

где α, β, γ и δ — числовые коэффициенты, значения которых и определяют специфику процедуры, ее нацеленность на решение той или иной экстремальной задачи.

Так, например, полагая $\alpha = \beta = -\delta = \frac{1}{2}$ и $\gamma = 0$, приходим к расстоянию, измеряемому по принципу «ближайшего соседа». Если же положить $\alpha = \beta = \delta = \frac{1}{2}$ и $\gamma = 0$, то расстояние между двумя классами определится как расстояние между двумя самыми далекими элементами этих классов, по принципу «дальнего соседа». И наконец, выбор коэффициентов соотношения (12.22) по формулам

$$\alpha = \frac{n_m}{n_m + n_q}, \quad \beta = \frac{n_q}{n_m + n_q}, \quad \gamma = \delta = 0$$

¹ См.: Keller W. J. Statistical via Personal Computers. «Compstat-86», Proceedings in Computational Statistics. — Wien, Physica-Verlag, 1986, pp. 332–337.

приводит к расстоянию $\rho_{\text{ср}}$ между классами, вычисленному как среднее из расстояний между всеми парами элементов, один из которых берется из одного класса, а другой — из другого.

Расстояния между классами в теоретико-вероятностной схеме кластер-анализа. В данной схеме анализируемая генеральная совокупность интерпретируется как *смесь унимодальных генеральных совокупностей*, каждая из которых и представляет один из искомых классов. Если дополнительно ограничить себя рассмотрением *только нормальных* классов, то каждый (l -й) такой класс, как известно (см. п. 3.1.5, соотношение (3.10)), определяется вектором средних значений $\mathbf{a}(l)$ и ковариационной матрицей $\Sigma(l)$. Для измерения расстояния между нормальными классами с номерами l и m используют так называемое «информационное расстояние Каллбэка» (S. Kullback)

$$\rho^2(S_l, S_m) = \frac{1}{2} (\mathbf{a}(l) - \mathbf{a}(m))^T (\Sigma^{-1}(l) + \Sigma^{-1}(m)) (\mathbf{a}(l) - \mathbf{a}(m)) + \frac{1}{2} \text{tr} \left\{ (\Sigma(l) - \Sigma(m)) (\Sigma^{-1}(l) - \Sigma^{-1}(m)) \right\}. \quad (12.23)$$

Если анализируемые классы (генеральные совокупности) различаются *только средними значениями* $\mathbf{a}(l)$ и $\mathbf{a}(m)$ (т.е. ковариационные матрицы у них одинаковы), то расстояние между ними измеряется так называемым «расстоянием Маталанобиса», которое вычисляется по формуле, получаемой из (12.23) при $\Sigma(l) = \Sigma(m) = \Sigma$:

$$\rho^2(S_l, S_m) = (\mathbf{a}(l) - \mathbf{a}(m))^T \Sigma^{-1} (\mathbf{a}(l) - \mathbf{a}(m)). \quad (12.24)$$

В статистической практике формулы (12.23) и (12.24) используются для вычисления расстояний между классами *и при отклонении распределения наблюдений внутри классов от нормального с заменой теоретических характеристик* $\mathbf{a}(j)$ и $\Sigma(j)$ *из оценками* $\hat{\mathbf{a}}(j)$ и $\hat{\Sigma}(j)$, построенными по наблюдениям, составляющим класс с номером j ($j = l, m$), см. выше формулы (12.11) и (12.12).

12.4.4. Функционалы качества разбиения на классы и экстремальная постановка задачи кластер-анализа

Естественно попытаться определить сравнительное качество различных способов разбиения заданной совокупности элементов на классы, т.е. определить тот количественный критерий, следуя которому можно было бы предпочесть одно разбиение другому. С этой целью в постановку

задачи кластер-анализа часто вводится понятие так называемого *функционала качества разбиения* $Q(S)$, определенного на множестве всех возможных разбиений. Функционалом он называется потому, что чаще всего разбиение S задается, вообще говоря, набором дискриминантных функций $\delta_1(X), \delta_2(X), \dots$. Тогда под наилучшим разбиением S^* понимается то разбиение, на котором достигается экстремум выбранного функционала качества. Выбор того или иного функционала качества, как правило, осуществляется весьма произвольно и опирается скорее на эмпирические и профессионально-интуитивные соображения, чем на какую-либо строгую формализованную систему.

Приведем примеры наиболее распространенных функционалов качества разбиения.

Функционалы качества разбиения при заданном числе классов. Пусть исследователем уже выбрана метрика d в пространстве $\Pi^p(X)$ и пусть $S = (S_1, S_2, \dots, S_k)$ — некоторое фиксированное разбиение наблюдений $X_1, X_2, \dots, X_n$ на заданное число k классов $S_1, S_2, \dots, S_k$.

За функционалы качества часто берутся следующие характеристики: *сумма («взвешенная») внутриклассовых дисперсий*

$$Q_1(S) = \sum_{l=1}^k \sum_{X_i \in S_l} d^2(X_i, \bar{X}(l)) \quad (12.25)$$

весьма широко используется в задачах кластер-анализа в качестве критерийной оценки разбиения;

сумма попарных внутриклассовых расстояний между элементами

$$Q_2(S) = \sum_{l=1}^k \sum_{X_i, X_j \in S_l} d^2(X_i, X_j)$$

либо

$$Q'_2(s) = \sum_{l=1}^k \frac{1}{n_l} \sum_{X_i, X_j \in S_l} d^2(X_i, X_j)$$

в большинстве ситуаций приводит к тем же наилучшим разбиениям, что и $Q_1(S)$, и тоже используется для сравнения кластер-процедур;

обобщенная внутриклассовая дисперсия $Q_3(S)$ является, как известно (см. п. 2.6.6, формула (2.30)), одной из характеристик степени рассеивания многомерных наблюдений одного класса (генеральной совокупности) около своего «центра тяжести». Следуя обычным правилам вычисления

выборочной ковариационной матрицы $\hat{\Sigma}_l$, отдельно по наблюдениям, попавшим в какой-то один класс S_l , получаем

$$Q_3(S) = \det \left(\sum_{l=1}^k n_l \hat{\Sigma}_l \right), \quad (12.26)$$

где под $\det A$ понимается «определитель матрицы A », а элементы $\hat{\sigma}_{qt}(l)$ выборочной ковариационной матрицы $\hat{\Sigma}_l$ класса S_l подсчитываются по формуле

$$\hat{\sigma}_{qt}(l) = \frac{1}{n_l} \sum_{X_i \in S_l} (x_i^{(q)} - \bar{x}^{(q)}(l))(x_i^{(t)} - \bar{x}^{(t)}(l)), \quad q, t = 1, 2, \dots, p, \quad (12.27)$$

где $x_i^{(v)}$ — v -я компонента многомерного наблюдения X_i , а $\bar{x}^{(v)}(l)$ — среднее значение v -й компоненты, подсчитанное по наблюдениям l -го класса.

Встречается и другой вариант использования понятия обобщенной дисперсии как характеристики качества разбиения, в котором операция суммирования $\hat{\Sigma}_l$ по классам заменена операцией умножения

$$Q_4(S) = \prod_{l=1}^k (\det \hat{\Sigma}_l)^{n_l}.$$

Как видно из формул (12.26) и (12.27), функционал $Q_3(S)$ является средней арифметической (по всем классам) характеристикой обобщенной внутриклассовой дисперсии, в то время как функционал $Q_4(S)$ пропорционален средней геометрической характеристике тех же величин.

Использование функционалов $Q_3(S)$ и $Q_4(S)$ особенно уместно в ситуациях, в которых исследователь в первую очередь задается вопросом: не сосредоточены ли наблюдения, разбитые на классы $S_2, S_2, \dots, S_k$, в пространстве размерности, меньшей, чем p ?

Функционалы качества разбиения при неизвестном числе классов. В ситуациях, когда исследователю заранее не известно, на какое число классов подразделяются исходные многомерные наблюдения $X_1, X_2, \dots, X_n$, функционалы качества разбиения $Q(S)$ выбирают чаще всего в виде простой алгебраической комбинации (суммы, разности, произведения, отношения) двух функционалов $I_1(S)$ и $I_2(S)$, один из которых I_1 является убывающей (невозрастающей) функцией числа классов k и характеризует, как правило, внутриклассовый разброс наблюдений, а второй I_2 — возрастающей (неубывающей) функцией числа классов k . При этом интерпретация функционала I_2 может быть различной. Под

I_2 понимается иногда и некоторая мера взаимной удаленности (близости) классов, и мера тех потерь, которые приходится нести исследователю при излишней детализации рассматриваемого массива исходных наблюдений, и величина, обратная так называемой «мере концентрации» всей структуры точек, полученной при разбиении исследуемого множества наблюдений на k классов.

Весьма гибким и достаточно общим подходом, реализующим идею одновременного учета двух функционалов, является подход, основанный на схеме, предложенной А. Н. Колмогоровым. Эта схема опирается на понятия *меры концентрации* $Z_\tau(S)$ точек, соответствующей разбиению S , и *средней меры внутриклассового рассеяния* $I_\tau^{(K)}(S)$, характеризующей то же разбиение S .

Под мерой концентрации $Z_\tau(S)$ предлагается понимать степенное среднее вида

$$Z_\tau(S) = M_\tau \left(\frac{v(X_1)}{n} \frac{v(X_2)}{n}, \dots, \frac{v(X_n)}{n} \right) = \left[\frac{1}{n} \sum_{i=1}^n \left(\frac{v(X_i)}{n} \right)^\tau \right]^{\frac{1}{\tau}}, \quad (12.28)$$

где $v(X_i)$ — число элементов в кластере, содержащем точку X_i , а выбор числового параметра τ находится в распоряжении исследователя и зависит от конкретных целей разбиения. При выборе τ полезно иметь в виду следующие частные случаи $Z_\tau(S)$:

$$Z_{-1}(S) = \frac{1}{k},$$

где k — число различных кластеров в разбиении S ; $\log Z_0(S) = \sum_{i=1}^k \frac{n_i}{n} \log \frac{n_i}{n}$ — естественная информационная мера концентрации (здесь n_i , как и ранее, число элементов в кластере S_i);

$$Z_\infty(S) = \max_{1 \leq i \leq k} \left(\frac{n_i}{n} \right);$$

$$Z_{-\infty}(S) = \min_{1 \leq i \leq k} \left(\frac{n_i}{n} \right);$$

$$Z_1(S) = \frac{1}{n} \sum_{j=1}^n \left(\frac{v(X_j)}{n} \right) = \frac{1}{n^2} \sum_{i=1}^k n_i^2.$$

Заметим, что при любом τ предложенная мера концентрации имеет минимальное значение, равное $1/n$, при разбиении исследуемого множества

на n одноточечных кластеров, и максимальное значение, равное 1, при объединении всех исходных наблюдений в один общий кластер.

При конструировании и сравнении различных кластер-процедур полезно иметь в виду, что объединение двух кластеров S_l и S_m в один дает прирост меры концентрации $Z_1(S)$, равный

$$\Delta Z_1 = \frac{1}{n^2} [(n_l + n_m)^2 - n_l^2 - n_m^2] = \frac{2n_l n_m}{n^2}.$$

Определение *средней меры внутриклассового рассеяния* $I_\tau^{(K)}(S)$ также опирается на понятие степенного среднего. В частности, полагают

$$I_\tau^{(K)}(S) = \left\{ \frac{1}{n} \sum_{i=1}^k n_i [Q_\tau^{(K)}(S_i)]^\tau \right\}^{\frac{1}{\tau}}, \quad (12.29)$$

где под

$$Q_\tau^{(K)}(S_i) = \left[\frac{1}{n_i} \sum_{X_j \in S_i} \sum_{X_l \in S_i} d^\tau(X_j, X_l) \right]^{\frac{1}{\tau}}$$

понимается обобщенная средняя мера рассеяния, характеризующая класс S_i . Числовой параметр τ здесь, как и прежде, выбирается по усмотрению исследователя. Полагая

$$Q_\tau^{(K)}(X) = \left[\frac{1}{v(X)} \sum_{X_l \in S(X)} d^\tau(X, X_l) \right]^{\frac{1}{\tau}},$$

где $S(X)$ — кластер, в который входит наблюдение X , а $v(X)$ — число элементов в кластере $S(X)$, формулу (12.29) можно переписать в виде

$$\begin{aligned} I_\tau^{(K)}(S) &= M_\tau(Q_\tau^{(K)}(X_1), \dots, Q_\tau^{(K)}(X_n)) \\ &= \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{v(X_i)} \sum_{X_l \in S(X_i)} d^\tau(X_i, X_l) \right]^{\frac{1}{\tau}}. \end{aligned} \quad (12.30)$$

При конструировании и сравнении различных кластер-процедур полезно иметь в виду, что объединение двух кластеров S_l и S_m в один дает прирост величины $n[I_\tau^{(K)}(S)]^\tau$, непосредственно характеризующей среднюю меру внутриклассового рассеяния, равный

$$\Delta [n(I_\tau^{(K)})^\tau] = \frac{n_l n_m}{n_l + n_m} \left\{ 2[\rho_\tau^{(K)}(S_l, S_m)]^\tau - [Q_\tau^{(K)}(S_l)]^\tau - [Q_\tau^{(K)}(S_m)]^\tau \right\}.$$

Очевидно, если ориентироваться на сокращение числа кластеров при наименьших потерях в отношении внутриклассового рассеивания, не обращая внимания на меру концентрации, то естественно объединять два кластера, для которых минимальна величина $\Delta(n[I_\tau^{(K)}]^\tau)$. Если же одновременно ориентироваться и на рост взвешенной концентрации $Z_1(S)$, то объединение кластеров следует подчинить требованию минимизации величины

$$\frac{\Delta[n(I_\tau^{(K)})^\tau]}{\Delta Z_1(S)} = \frac{\{2[\rho_\tau^{(K)}(S_l, S_m)]^\tau - [Q_\tau^{(K)}(S_l)]^\tau - |Q_\tau^{(K)}(S_m)|^\tau\} n^2}{n_l + n_m}.$$

12.4.5. Формулировка экстремальных задач разбиения исходного множества объектов на классы при неизвестном числе классов

Возможно множество вариантов таких формулировок. Рассмотрим два из них, относящихся к наиболее естественным и часто используемым.

В а р и а н т 1: комбинирование функционалов качества. Требуется найти такое разбиение S^* , для которого некоторая алгебраическая комбинация функционала, характеризующего среднее внутриклассовое рассеивание (12.30), и функционала, характеризующего меру концентрации полученной структуры (12.28), достигала бы своего экстремума. В качестве примеров можно привести комбинации, задаваемые формулами вида

$$\alpha I_1(S) + \beta I_2(S) \quad \text{и} \quad [I_1(S)]^\alpha [I_2(S)]^\beta, \quad (12.31)$$

где $I_1(S) = I_\tau^{(K)}(S)$; $I_2(S) = \frac{1}{Z_\tau(S)}$; α и β — некоторые положительные константы, например, $\alpha = \beta = 1$.

В а р и а н т 2: двойственная формулировка. Требуется найти разбиение S^* , которое, обладая концентрацией $Z_\tau(S^*)$, не меньшей заданного порогового значения Z_0 , давало бы наименьшее внутриклассовое рассеивание $I_\tau^{(K)}(S^*)$, или, в двойственной подстановке: при заданном пороговом значении I_0 найти разбиение S^* с внутриклассовым рассеиванием $I_\tau^{(K)}(S^*) \leq I_0$ и наибольшей концентрацией $Z_\tau(S^*)$.

12.4.6. Основные типы задач кластер-анализа и основные типы кластер-процедур

Прежде всего целесообразно подразделение всех задач кластер-анализа на два основных типа: B_1 и B_2 в зависимости от объема n совокупности классифицируемых наблюдений $X_1, X_2, \dots, X_n$.

К типу B_1 отнесем задачи классификации сравнительно небольших по объему совокупностей наблюдений, состоящих, как правило, не более чем из нескольких десятков наблюдений. Сюда, по-видимому, могут быть отнесены задачи классификации некоторых макрообъектов, таких, как страны, города, фирмы, предприятия, типы технологических процессов и т. п.

К типу B_2 будем относить задачи классификации достаточно больших массивов многомерных наблюдений (n — порядка нескольких сотен и тысяч; классификация индивидуумов, семей, изделий, некоторых промышленных и технических микрообъектов). Подобное разделение задач классификации на два типа хотя и условно, но весьма необходимо, и в первую очередь с точки зрения принципиального различия идей и методов, на основании которых конструируются кластер-процедуры в том и в другом случае. Например, для задач типа B_2 целесообразно построение процедур последовательного типа, обладающих достаточно хорошими, хотя бы асимптотическими по n , свойствами.

С точки зрения *априорной информации об окончательном числе классов*, на которое требуется разбить исследуемую совокупность объектов, задачи кластер-анализа можно подразделить на три основных типа:

- (а) число классов априори задано;
- (б) число классов неизвестно и подлежит определению (оценке);
- (в) число классов неизвестно, но его определение и не входит в условие задачи; требуется построить так называемое *иерархическое дерево* исследуемой совокупности, или *дендрограмму*.

В соответствии с подразделением задач кластер-анализа на типы можно выделить следующие три основных типа обслуживающих их кластер-процедур:

- *процедуры иерархические* (агломеративные и дивизимные). Предназначены в основном для решения задач типа (в). Что касается объема классифицируемой совокупности, то формально иерархические процедуры применимы и для задач B_1 , и для задач B_2 . Однако поскольку эти процедуры основаны на переборе элементов матрицы расстояний $\rho(X_i, X_j)$ (или матрицы соответствующих мер близости), то конструктивно реализуемыми их можно признать лишь в пределах задач типа B_1 . Следует отметить, что иерархические процедуры применяются иногда и для решения задач типов B_1 , (а) и B_1 , (б);

- *процедуры параллельные*. Предназначены для решения задач типов B_1 , (а) и B_1 , (б). Они реализуются с помощью итерационных алгоритмов, на каждом шаге которых одновременно (параллельно) используются все имеющиеся у нас наблюдения;

- *процедуры последовательные*. Предназначены в основном для реше-

ния задач типов $B_2, (a)$ и $B_2, (b)$. Они реализуются с помощью итерационных алгоритмов, на каждом шаге которых используется лишь небольшая часть, например одно из исходных наблюдений, а также результат разбиения на предыдущем шаге.

12.4.7. Иерархические процедуры

Как отмечалось выше, принцип работы иерархических агломеративных (дивизимных) процедур состоит в последовательном объединении (разделении) групп элементов сначала самых близких (далеких), а затем все более отдаленных друг от друга (приближенных друг к другу). При этом агломеративные процедуры начинают обычно с объединения отдельных элементов, а дивизимные — с разъединения всей исходной совокупности наблюдений.

С некоторой точки зрения иерархические процедуры, по сравнению с другими кластер-процедурами, дают более полный и тонкий анализ структуры исследуемого множества наблюдений. Привлекательной стороной подобных алгоритмов является и возможность наглядной интерпретации проведенного анализа. Легко себе представить также использование иерархических процедур и для решения задач кластер-анализа типов (а) и (б), т. е. для разбиения наблюдений на какое-то объективно обусловленное число классов, заданное или неизвестное. При решении задач типа (а) для этого, очевидно, следует продолжать реализацию иерархического алгоритма до тех пор, пока число различных классов не станет равным априори заданному числу k . При решении задач типа (б) естественно было бы подчинить правило остановки иерархической процедуры одному из критериев качества разбиения.

К недостаткам иерархических процедур следует отнести громоздкость их вычислительной реализации. Соответствующие алгоритмы на каждом шаге требуют вычисления всей матрицы расстояний, а следовательно, емкой машинной памяти и большого времени. Поэтому реализация таких алгоритмов при числе наблюдений, большем нескольких сотен, оказывается либо невозможной, либо нецелесообразной.

Кроме того, имеется широкий класс достаточно естественных примеров, в которых иерархические процедуры, даже подчиненные на каждом шаге некоторому критерию качества разбиения, приводят для любого наперед заданного числа кластеров k к разбиению, весьма далекому от оптимального в смысле того же самого критерия качества. Если прибавить к этому широкое экспериментальное подтверждение того же эффекта, то можно прийти к выводу, что «конечная неоптимальность» оптималь-

ного иерархического алгоритма является скорее правилом, чем исключением. Специфический характер метода образования групп, свойственный иерархическим процедурам, оказывается, по-видимому, слишком жестким ограничением с точки зрения экстремального подхода к решению задач классификации наблюдений при определенном числе классов.

Приведем некоторые примеры иерархических алгоритмов:

- *агломеративный иерархический алгоритм «ближайшего соседа»* (или «одной связи»). Этот алгоритм исходит из матрицы расстояний между наблюдениями, в которой расстояние между кластерами определено по правилу «ближайшего соседа» (см. выше). На первом шаге алгоритма каждое наблюдение X_i ($i = 1, 2, \dots, n$) рассматривается как отдельный кластер. Далее на каждом шаге работы алгоритма происходит объединение двух самых близких кластеров и соответственно пересчитывается матрица расстояний, размерность которой, естественно, снижается на единицу. Работа алгоритма заканчивается, когда все исходные наблюдения объединены в один класс. Поскольку расстояние между любыми двумя кластерами в этом алгоритме равно расстоянию между двумя самыми близкими элементами, представляющими свои классы, то получаемые в итоге кластеры могут иметь достаточно сложную форму, в частности, они не обязаны быть выпуклыми; ведь два элемента (наблюдения) попадают в один кластер, если существует соединяющая их цепочка близких между собой элементов (так называемый «цепочечный эффект»). Это обстоятельство можно отнести как к достоинствам алгоритма, так и к его недостаткам.

Существуют различные способы устранения *цепочечного эффекта* при образовании классов с помощью алгоритма ближайшего соседа. Наиболее простым и естественным из них можно признать, например, введение ограничения сверху на максимальное расстояние между элементами одного класса: если при формировании классов для некоторых элементов получаемого кластера взаимное расстояние превысит некоторый заданный порог, то эти элементы следует разнести по какому-то дополнительному правилу в разные классы.

- *агломеративные иерархические алгоритмы «средней связи» и «полной связи»* (или «дальнего соседа»). Эти алгоритмы отличаются от описанного выше алгоритма «ближайшего соседа» лишь способом вычисления расстояния между классами. В алгоритме средней связи под расстоянием между кластерами понимается среднее из расстояний между всевозможными парами представителей этих кластеров. В алгоритме полной связи (или дальнего соседа) расстояние между двумя кластерами определяется как расстояние между двумя самыми отдаленными друг от

друга представителями своих кластеров.

- *К-обобщенная иерархическая процедура*, т.е. обобщенная по Колмогорову. Поскольку все вышеперечисленные виды расстояний между кластерами могут быть получены в качестве частных случаев обобщенного расстояния Колмогорова (12.21), то представляется естественным ввести понятие *К-обобщенной иерархической процедуры*. Очевидно, в класс *К-обобщенных иерархических процедур* следует включить все обычные иерархические алгоритмы, использующие в качестве расстояний между кластерами обобщенное расстояние Колмогорова (12.21) при том или другом конкретном выборе числового параметра τ ;

- *процедуры иерархические, использующие понятие порога*. Общая схема подобных процедур отличается от обычной логической схемы ранее описанных иерархических процедур лишь дополнительным заданием последовательности, как правило, монотонной, порогов $c_1, c_2, \dots, c_t$, которые используются следующим образом. Для определенности дадим пояснения для агломеративных процедур. На первом шаге алгоритма попарно объединяются элементы, расстояние между которыми не превосходит величины c_1 , либо мера близости которых не менее c_1 . На втором шаге алгоритма объединяются элементы, или группы элементов, расстояние между которыми не превосходит c_2 , либо мера близости которых не менее c_2 , и т.д. Очевидно, при $c_t = \infty$ или, при сравнении мер близости, при $c_t = 0$, на последнем t -м шаге все элементы исходной совокупности окажутся объединенными в один общий класс. Заметим, однако, что объединение в кластеры, подчиненные подобным пороговым иерархическим алгоритмам, приводит к образованию, вообще говоря, пересекающихся промежуточных классов, которые могут не расцепиться вплоть до последнего шага. Поэтому эффективность подобных процедур, возможность выбора подходящих пороговых значений $c_1, \dots, c_t$ существенно зависят от внутренней геометрической структуры исходного множества наблюдений. В частности, пороговые иерархические процедуры оказываются уместными и достаточно эффективными в ситуациях, когда отсутствует (или слабо выражен) цепочечный эффект в структуре исходной совокупности наблюдений и когда последние, естественно, распадаются на какое-то количество достаточно отдаленных друг от друга отдельных скоплений точек в исследуемом факторном пространстве.

12.4.8. Параллельные кластер-процедуры

В алгоритмах кластер-анализа реализуется обычно одна из двух основных родственных идей, которой исследователь хочет подчинить свое

разбиение на классы. Это либо идея оптимизации разбиения в смысле заранее выбранного функционала качества разбиения, либо идея образования кластеров по принципу определения мест наибольшей сгущенности (плотности, концентрации) точек наблюдений в рассматриваемом факторном пространстве. Коль скоро характер параллельных процедур предусматривает одновременный обсчет всех исходных наблюдений на каждом шаге алгоритма, то естественно попытаться решать поставленную задачу с помощью обычного перебора различных вариантов разбиения.

Однако нетрудно подсчитать, что уже при сравнительно небольшом общем числе классифицируемых точек n (порядка нескольких десятков) полный перебор всех вариантов разбиения на заданное, а тем более на неизвестное число классов практически неосуществим. Следовательно, основной смысл конструирования различных параллельных алгоритмов классификации — в указании способа сокращения числа перебора вариантов, в описании пути, приводящего, быть может, лишь к приближенному решению поставленной задачи, но пути, конструктивно реализуемого и не слишком дорогого.

а) *Алгоритмы, связанные с функционалами качества разбиения.* К таким алгоритмам следует в первую очередь отнести алгоритмы «последовательного переноса точек из класса в класс». Эти алгоритмы отправляются от некоторого начального разбиения $S^0 = \{S_1^{(0)}, \dots, S_k^{(0)}\}$, полученного произвольно или с помощью какого-либо из методов предварительной обработки исходных наблюдений. Вычисляется значение принятого критерия качества разбиения $Q(S)$, например, вида (12.25) при заданном числе классов k или вида (12.31) при неизвестном числе классов. Затем каждое из наблюдений X_i поочередно перемещается во все кластеры (при этом оно рассматривается как самостоятельный кластер, если число кластеров неизвестно) и оставляется в том положении, которое соответствует наилучшему значению функционала качества Q . Работа алгоритма заканчивается, когда перемещения наблюдений перестанут приводить к улучшению (в смысле Q) качества разбиения. Часто описанный алгоритм применяют несколько раз к одной и той же исходной совокупности наблюдений, начиная с разных начальных разбиений $S^{(0)}$, и выбирают в итоге наилучший (в смысле Q) вариант разбиения.

б) *Алгоритмы, использующие понятие эталонных точек (множеств).* Опишем общую формальную схему одного достаточного широкого класса алгоритмов, реализация которой может приводить как к параллельным, так и к последовательным кластер-процедурам.

Под эталонными множествами $E_1, E_2, \dots, E_k$ будем понимать каким-то образом, в частности случайным, сформированные непересекающиеся

подмножества исходной совокупности наблюдений $\{X_1, X_2, \dots, X_n\}$ заранее определенных объемов соответственно $m_1, m_2, \dots, m_k$. Как правило, $m_1 + m_2 + \dots + m_k$ составляет лишь незначительную долю от общего числа исходных наблюдений n . В частном случае $m_1 = m_2 = \dots = m_k = 1$ будем иметь набор k эталонных точек. Пусть для любого набора эталонных множеств $E = \{E_1, E_2, \dots, E_k\}$, для любого наблюдения X_i , для любой группы исходных наблюдений A и для произвольного $l = 1, 2, \dots, k$ заданы некоторые специальные функции $\varphi(X_i, A)$ и $\psi(X_i, S_l, E)$, характеризующие меру типичности точки X_i как представителя группы точек A (меру однородности наблюдений X_i и группы наблюдений A) и меру типичности точки X_i как представителя класса S_l , построенного с использованием эталонного множества E_l из E .

Будем для определенности считать, что чем меньше значения функций φ и ψ , тем типичнее соответствующая точка в указанном выше смысле. Тогда общая схема эталонных алгоритмов может быть представлена следующим образом. При заранее заданном числе классов k каким-то образом (случайным, с помощью обучающих выборок, из экспертно-профессиональных соображений или с помощью методов предварительной обработки исходных данных) выбираются числа $m_1, m_2, \dots, m_k$ и начальная система эталонных множеств $E^{(0)} = \{E_1^{(0)}, E_2^{(0)}, \dots, E_k^{(0)}\}$. Класс $S_l^{(0)}$ формируется из наблюдений, наиболее типичных с точки зрения представительства эталона $E_l^{(0)}$, т. е.

$$S_l^{(0)} = \{X_i: \varphi(X_i, E_l^{(0)}) < \varphi(X_i, E_j^{(0)}), \quad j = 1, 2, \dots, k, \quad j \neq l\}.$$

Если оказывается, что $\varphi(X_i, E_l^{(0)}) = \varphi(X_i, E_j^{(0)})$, то можно условиться относить наблюдение X_i к тому из классов $S_l^{(0)}$ и $S_j^{(0)}$, который обладает меньшим порядковым номером. Затем строится новая система эталонных множеств $E = \{E_1^{(1)}, E_2^{(2)}, \dots, E_k^{(1)}\}$, в которой эталон $E_l^{(1)}$ формируется из m_l точек, дающих m_l наименьших значений функции $\psi(X, S_l^{(0)}, E^{(0)})$. После этого по тому же правилу строят новое разбиение $S^{(1)} = \{S_1^{(1)}, S_2^{(1)}, \dots, S_k^{(1)}\}$, но уже относительно эталонов $E^{(1)}$ и т. д. Итерации продолжают до тех пор, пока не получают устойчивых классов, т. е. до такого номера v , при котором $S^{(v)} = S^{(v+1)}$.

Если число классов, на которое требуется разбить исходную совокупность наблюдений, *заранее не известно*, то в описанную схему необходимо ввести некоторые дополнения. В частности, на начальном этапе устанавливаются: числа $k^{(0)}, m_1, m_2, \dots, m_{k^{(0)}}$, система эталонных множеств $E^{(0)} = \{E_1^{(0)}, E_2^{(0)}, \dots, E_{k^{(0)}}^{(0)}\}$, а также величина φ_0 — минимально

возможная типичность точки, представляющей свой класс, и ψ_0 — минимально возможная нетипичность двух разных классов¹. Как и прежде, вначале подсчитываются $\varphi(X_i, E_l^{(0)})$ для всех наблюдений ($i = 1, 2, \dots, n$) и для всех эталонов ($l = 1, 2, \dots, k^{(0)}$). Последовательно для каждого i определяются эталоны $E_{l(i)}^{(0)}$, для которых данное i -е наблюдение является наиболее типичным, т. е.

$$\varphi(X_i, E_{l(i)}^{(0)}) = \min_{1 \leq l \leq k^{(0)}} \varphi(X_i, E_l^{(0)}).$$

Если $\varphi(X_1, E_{l(1)}^{(0)}) \leq \varphi_0$, то наблюдение X_1 включается в состав класса $S_{l(1)}^{(0)}$; если же $\varphi(X_1, E_{l(1)}^{(0)}) > \varphi_0$, то наблюдение X_1 принимается одновременно за новый эталон $E_{k^{(0)}+1}^{(0)}$ и за новый класс $S_{k^{(0)}+1}^{(0)}$. Затем та же процедура сравнения производится с $\varphi(X_2, E_{l(2)}^{(0)})$ и т. д. до $\varphi(X_n, E_{l(n)}^{(0)})$. Пусть в результате у нас образовалось на этом этапе алгоритма $k^{(0)}(0)$ классов и столько же эталонов $\tilde{E} = \{E_1^{(0)}, \dots, E_{k^{(0)}(0)}^{(0)}\}$ (очевидно $k^{(0)}(0) \geq k^{(0)}$). После этого полученные классы $S_1^{(0)}, S_2^{(0)}, \dots, S_{k^{(0)}(0)}^{(0)}$ проверяются на значимое попарное различие с помощью порога ψ_0 , а именно, если

$$\begin{aligned} \psi_{q,l}^{(0)} &= \psi(S_q^{(0)}, S_l^{(0)}) \\ &= \frac{1}{2} \left[\frac{1}{n_q} \sum_{X_i \in S_q} \psi(X_i, S_l, E^{(0)}) + \frac{1}{n_l} \sum_{X_j \in S_l} \psi(X_j, S_q, E^{(0)}) \right] < \psi_0, \end{aligned}$$

то классы $S_q^{(0)}$ и $S_l^{(0)}$ объединяются в один класс $S_{(q,l)}$ с номером $z(q, l)$, равным $\min(q, l)$, и с эталоном $E_{z(q,l)}^{(0)}$, состоящим из $\max(m_q, m_l)$ наблюдений, дающих соответствующее число наименьших значений функции $\psi(X, S_{z(q,l)}, E^{(0)})$. Последовательное вычисление величин $\psi(S_q^{(0)}, S_l^{(0)})$ и их сравнение с ψ_0 производится до тех пор, пока не окажется, что

$$\min_{q,l} \psi_{q,l}^{(0)} \geq \psi_0.$$

Это означает окончание первой итерации алгоритма и образование новой системы эталонов $E^{(1)} = \{E_1^{(1)}, \dots, E_{k^{(1)}}^{(1)}\}$ и соответствующих *значимо*

¹ В наиболее общих процедурах «пороги» φ_0 и ψ_0 , так же как и общее число $n^{(v)}$ элементов ($n^{(v)} \leq n$), составляющих классы $S_1^{(v)}, \dots, S_{k^{(v)}}^{(v)}$, задаются переменными, т. е. изменяющимися по определенному правилу при переходе от одного этапа алгоритма к другому.

различимых классов $S^{(0)} = \{S_1^{(0)}, S_2^{(0)}, \dots, S_{k(1)}^{(0)}\}$. Затем процедура повторяется применительно к эталонам $E^{(1)}$ и т. д. до получения устойчивого разбиения $S^{(v)}$ всех исходных элементов на некоторое число классов $k^{(v+1)}$.

При конструировании функций $\varphi(X, A)$ и $\psi(X, S_l, E)$ можно использовать заданную метрику (расстояние между точками) $\rho(X, Y)$ или меру близости $r(X, Y)$, а также подходящим образом выбранный вариант обобщенного расстояния типа (12.21).

Используются, в частности, следующие способы задания φ и ψ :

$$\begin{aligned}\varphi(X, A) &= \sum_{X_i \in A} \rho(X, X_i), \\ \varphi(X, A) &= \rho(X, \bar{X}(A)),\end{aligned}$$

где $\bar{X}(A)$ — одна из разновидностей среднего значения всех наблюдений X_i , принадлежащих группе A . Например, $\bar{X}(A)$ может быть обычным арифметическим средним, т. е. «центром тяжести» группы A ;

$$\begin{aligned}\psi(X, S_l, E) &= \frac{\varphi(X, S_l) \cdot \varphi(X, E_l)}{\left[\sum_{j=1}^k \varphi(X, E_j) \right]^2}, \\ \psi(X, S_l, E) &= \varphi(X, S_l).\end{aligned}$$

В качестве примера *параллельных кластер-процедур, укладываемых в описанную общую схему*, рассмотрим теперь серию алгоритмов, объединенных названием «Форель». Общую для этих алгоритмов идею проиллюстрируем на примере алгоритма «Форель-1».

Пусть совокупность $\{X_1, X_2, \dots, X_n\}$ нужно разбить каким-то образом на некоторое число классов (заранее не известное). Пусть $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ и R_0 — радиус минимальной гиперсферы с центром в $\bar{X}$, содержащей все точки исследуемой совокупности. Зададим произвольный радиус $R < R_0$ и рассмотрим процедуру выделения классов для заданного R . Из любой точки $X_i = \bar{X}_1$, принятой за центр, радиусом R описывается гиперсфера C_1 . Находится центр тяжести $\bar{X}_2$ точек совокупности, попавших в C_1 . Из $\bar{X}_2$ радиусом R описывается гиперсфера C_2 и определяется $\bar{X}_3$ — центр тяжести точек исследуемой совокупности, попавших в C_2 . Процедура построения гиперсфер и точек $\bar{X}_k$ повторяется до тех пор, пока точки $\bar{X}_k$ не перестанут меняться. Точки совокупности, попавшие в «остановившуюся» гиперсферу, принимаются за первый класс S_1 . Для всех оставшихся точек, т. е. не попавших в класс S_1 , вновь применяется описанная выше

процедура, выделяющая еще один класс S_2 , и т. д. до тех пор, пока все точки совокупности не будут распределены по классам.

Применение описанного алгоритма для ряда последовательных значений $R^{(v)} = R_0 - v\Delta$, ($\Delta = \frac{R_0}{N}$, $v = 1, 2, \dots, N - 1$) позволяет ориентировочно оценить наиболее предпочтительное число классов для данной совокупности объектов. При этом основанием для выбора числа классов может служить многократное повторение одного и того же числа классов для нескольких последовательных значений $R^{(v)}$ и его резкое возрастание на следующем шаге.

Если ставится задача разбить совокупность на заданное число классов, то одна из моделей модификации алгоритма «Форель-1» — «Форель-2» методом последовательных приближений позволяет находить минимальный радиус $R_{\min}$, дающий разбиение на заданное число классов.

Покажем теперь, как процедура выделения одного класса точек из совокупности, являющаяся основой алгоритмов типа «Форель», может быть изложена в рамках описанной выше общей схемы «эталонных алгоритмов». Выделение класса точек из совокупности эквивалентно разделению совокупности на два класса. В качестве набора эталонных множеств $E = (E_1, E_2)$ будем брать две точки $E = (e_1, e_2)$. Функцию $\varphi(X, e_i)$ определим следующим образом: $\varphi(X, e_1) = \rho_E^2(X, e_1)$; $\varphi(X, e_2) = R$, где R — выбранная величина радиуса. Функция $\psi(X, S_l, E)$ может быть определена многими способами, например так:

$$\psi(X, S_l, E) = \sum_{X_i \in S_l} \rho_E^2(X, X_i).$$

12.4.9. Последовательные кластер-процедуры

Если число n классифицируемых наблюдений $X_1, X_2, \dots, X_n$ достаточно велико (от нескольких сотен и более), то, как мы уже отмечали, реализация кластер-процедур иерархического и параллельного типов крайне трудоемка, а иногда и практически невозможна. В этих случаях пользуются итерационными алгоритмами, на каждом шаге которых последовательно обсчитывается лишь *небольшая часть* исходных наблюдений, например *одно из них*. В том, что n велико, имеются не только неудобства, но и свои преимущества. В частности, это позволяет исследовать асимптотические (по n) свойства соответствующих процедур, аналогичные, например, свойствам состоятельности, асимптотической несмещенности и т. п., анализируемым в теории статистического оценивания и статистической проверки гипотез.

Как и в параллельных алгоритмах, основными средствами и идеями, при конструировании последовательных кластер-процедур являются: мера близости или расстояние между группами; порог; эталонные множества или точки; функционал качества разбиения.

Так же, как и прежде, более простой, а, главное, всегда имеющей решение, является обычная задача типизации, при которой исходное множество многомерных наблюдений разбивается на определенное число «областей группирования» по принципу наперед заданной взаимной близости элементов, отнесенных к одной области группирования. Простейшим примером такого рода является разбиение на интервалы группирования исходной выборки одномерных наблюдений, особенно необходимое как раз при достаточно больших объемах выборки n .

Именно такую задачу решает, например, простой последовательный алгоритм, использующий понятие порога c . В этом алгоритме случайным образом выбирается точка X_1 , которая объявляется центром e_1 первой группы. Затем точка X_2 относится к первой группе, если $\rho(X_2, e_1) \leq c$. В противном случае X_2 принимается за центр второй группы $X_2 = e_2$ и т. д. На l -м шаге, когда уже имеется r групп, точка X_l либо становится центром $(r - 1)$ -й группы, либо относится к той из групп, для которой $\rho(X_l, e_j) \leq c$. Если таких групп несколько, то выбирается та, к центру которой точка X_l ближе всего; если и таких групп несколько, то устанавливаются некоторые соглашения о том, куда относить X_l в этом случае.

Остановимся далее на описании двух наиболее общих и наиболее исследованных последовательных кластер-процедур.

а) Метод k -средних. Пусть наблюдения $X_1, X_2, \dots, X_n$ требуется разбить на заданное число k ($k \leq n$) однородных (в смысле некоторой метрики ρ) классов.

Смысл описываемого алгоритма — в последовательном уточнении эталонных точек $E^{(v)} = \{e_1^{(v)}, e_2^{(v)}, \dots, e_k^{(v)}\}$ (v — номер итерации, $v = 0, 1, 2, \dots$) с соответствующим пересчетом приписываемых им «весов» $\Omega^{(v)} = \{\omega_1^{(v)}, \omega_2^{(v)}, \dots, \omega_k^{(v)}\}$. При этом нулевое приближение $E^{(0)}$ строится с помощью случайно выбранных первых k точек исследуемой совокупности, т. е.

$$\begin{aligned} e_i^{(0)} &= X_i, \\ \omega_i^{(0)} &= 1, \quad i = 1, 2, \dots, k. \end{aligned}$$

Затем на l -м шаге «извлекается» точка X_{k+1} и выясняется, к какому из эталонов $e_i^{(0)}$ она оказалась ближе всего. Именно этот, самый близкий к X_{k+1} , эталон заменяется эталоном, определяемым как центр тяжести старого эталона и присоединенной к нему точки X_{k+1} (с увеличением на

единицу соответствующего ему веса), а все другие эталоны остаются неизменными (с прежними весами) и т. д. Таким образом, пересчет эталонов и весов на v -м шаге, т. е. при извлечении очередной точки X_{k+v} , происходит по следующему правилу:

$$e_i^{(v)} = \begin{cases} \frac{\omega_i^{(v-1)} e_i^{(v-1)} + X_{k+v}}{\omega_i^{(v-1)} + 1}, & \text{если } \rho(X_{k+v}, e_i^{(v-1)}) = \min_{1 \leq j \leq k} \rho(X_{k+v}, e_j^{(v-1)}), \\ e_i^{(v-1)} & \text{в противном случае,} \end{cases}$$

$$\omega_i^{(v)} = \begin{cases} \omega_i^{(v-1)} + 1, & \text{если } \rho(X_{k+v}, e_i^{(v-1)}) = \min_{1 \leq j \leq k} \rho(X_{k+v}, e_j^{(v-1)}), \\ \omega_i^{(v-1)}, & \text{в противном случае,} \end{cases}$$

$i = 1, 2, \dots, k.$

При этом если обнаруживается несколько (по i) одинаковых минимальных значений $\rho(X_{k+v}, e_i^{(v-1)})$, то можно условиться относить точку X_{k+v} к эталону с минимальным порядковым номером.

При достаточно большом числе итераций или при достаточно больших объемах классифицируемых совокупностей n и при весьма широких ограничениях на природу исследуемых наблюдений дальнейший пересчет эталонных точек практически не приводит к их изменению, т. е. имеет место сходимость (в определенном смысле) $\mathbf{E}^{(v)}$ к некоторому пределу при $v \rightarrow \infty$.

Если же в какой-то конкретной задаче исследователь не успел добраться до стадии практически устойчивых (по v) значений эталонных точек, то пользуются одним из двух вспомогательных приемов. Либо «защелкивают» алгоритм, «прогоняя» его после рассмотрения последней точки $X_n = X_{k+(n-k)}$ снова через точку X_1 , затем X_2 , и т. д., либо производят многократное повторение алгоритма, используя в качестве начального эталона $\mathbf{E}^{(0)}$ различные комбинации из k точек исследуемой совокупности и выбирая для дальнейшего наиболее повторяющийся (в некотором смысле) финальный эталон $\mathbf{E}^{(n-k)}$.

Окончательное разбиение S исследуемой совокупности многомерных наблюдений на k классов производится в соответствии с правилом описанного выше минимального дистанционного разбиения $S(\mathbf{E})$ относительно центров тяжести (эталонов) $\mathbf{E} = \mathbf{E}^{(n-k)}$, которое, кстати, является частным случаем разбиений ранее описанной общей схемы эталонных алгоритмов, получающихся при

$$\varphi(X, E_l) = \rho(X, E_l),$$

т. е.

$$S_l(E) = \{X: \rho(X, E_l) < \rho(X, E_j); j = 1, 2, \dots, k; j \neq l\}.$$

Если оказывается, что $\rho(X, E_l) = \rho(X, E_j)$, то точку X относят к тому из классов S_l и S_j , который обладает меньшим порядковым номером.

Проведенные исследования свойств метода k -средних (см., например, [Айвазян С. А. и др., 1989] говорят о том, что в достаточно общих ситуациях при больших объемах выборочных совокупностей этот алгоритм строит разбиение, близкое к наилучшему в смысле функционала (12.25).

б) Метод k -средних при неизвестном числе классов. Для обобщения описанного метода на случай *неизвестного* числа классов следует задаться двумя константами Φ_0 и Ψ_0 , названными соответственно *мерой грубости* и *мерой точности*. Работа алгоритма также состоит в последовательном построении эталонных точек $E^{(v)} = (E_1^{(v)}, \dots, E_{k(v)}^{(v)})$ и весов $\omega_1^{(v)}, \dots, \omega_{k(v)}^{(v)}$, но число классов $k(v)$ может меняться при этом от итерации к итерации.

На нулевом шаге итерации берется любое начальное $k(0)$ и полагается

$$\omega_j^{(0)} = 1, \quad E_j^{(0)} = X_j \quad (j = 1, \dots, k(0)).$$

Затем производится процедура «огрубления» эталонных точек. А именно, подсчитывается расстояние между двумя ближайшими эталонными точками и сравнивается это расстояние с заданной мерой грубости Φ_0 . Если это минимальное расстояние меньше Φ_0 , то соответствующая пара эталонных точек заменяется их взвешенным средним с весом, равным сумме соответствующих двух весов. Процедура огрубления заканчивается тогда, когда расстояние между любыми двумя эталонными точками не меньше чем Φ_0 . Пусть в результате процедуры огрубления мы имеем число эталонных точек $\bar{k}(0)$ ($\bar{k}(0) \leq k(0)$), эталонные точки $E_j^{(0)}$ ($j = 1, \dots, \bar{k}(0)$) и веса $\omega_j^{(0)}$ ($j = 1, \dots, \bar{k}(0)$).

На первом шаге итерации извлекается точка $X_{k(0)+1}$ и вычисляется расстояние от $X_{k(0)+1}$ до ближайшей к ней эталонной точки $E_j^{(0)}$ ($j = 1, \dots, \bar{k}(0)$). При этом если это расстояние больше Ψ_0 , то $X_{k(0)+1}$ объявляется новой эталонной точкой $E_{k(0)+1} = X_{k(0)+1}$ с весом $\omega_{k(0)+1} = 1$, а все остальные эталонные точки и соответствующие им веса остаются неизменными.

Если это минимальное расстояние меньше чем Ψ_0 , то самый близкий к $X_{k(0)+1}$ эталон заменяется эталоном, определяемым как центр тяжести старого эталона и присоединенной к нему точки $X_{k(0)+1}$. Вес точки

$X_{k(0)+1}$ считается равным 1. Вес этого нового эталона равен сумме весов объединяемых точек (старого эталона и точки $X_{k(0)+1}$).

Все остальные эталоны и соответствующие веса остаются неизменными. Таким образом, пересчет эталонов и весов в этом случае происходит точно так же, как и в обычном методе k -средних.

После процедуры огрубления эталонных точек переходят ко 2-му шагу итерации и так далее.

Выбирая различные константы Φ_0 , Ψ_0 , мы будем с помощью этого алгоритма получать различные разбиения. Выбор величин Φ_0 и Ψ_0 можно считать удачным, если разбиение, соответствующее этим величинам, признается оптимальным или с точки зрения экспертов, или в смысле принятых функционалов качества разбиения.

ВЫВОДЫ

1. Разделение рассматриваемой совокупности объектов или явлений на однородные (в определенном смысле) группы называется *классификацией*. При этом термин «классификация» используют, в зависимости от контекста, для обозначения как *самого процесса* разделения, так и *его результата*. Это понятие тесно связано с такими терминами, как группировка, типологизация, систематизация, дискриминация, кластеризация, и является одним из основополагающих в практической и научной деятельности человека.

2. Среди типов прикладных задач классификации следует выделить:

- 1) *комбинационные группировки* и их непрерывные обобщения — разбиение совокупности на интервалы (области) группирования;
- 2) *простая типологизация*: выявление естественного расслоения анализируемых данных (объектов) на четко выраженные «сгустки» (кластеры), лежащие друг от друга на некотором расстоянии, но не разбивающиеся на столь же удаленные друг от друга части;
- 3) *связная неупорядоченная типологизация*: использование реализованной в пространстве *результатирующих* показателей простой типологизации в качестве обучающих выборок при классификации той же совокупности объектов в пространстве *описательных* признаков;
- 4) *связная упорядоченная типологизация*, которая отличается от связной неупорядоченной возможностью *экспертного упорядочения* классов, полученных в пространстве *результатирующих* показателей, и использованием этого упорядочения для построения сводного латентного *результатирующего* показателя как функции от описательных переменных;
- 5) *структурная типологизация* дает на «выходе» задачи дополнительно к описанию классов еще и описание существующих между ними и их элементами структурных (в том числе иерархических) связей;
- 6) *типоло-*

гизация динамических траекторий системы: в качестве классифицируемых объектов выступают характеристики динамики исследуемых систем, например, дискретные или непрерывные временные ряды или траектории систем, которые в каждый момент времени могут находиться в одном из заданных состояний.

3. Определяющим моментом в выборе *подходящей математической постановки* для конкретной задачи классификации является ответ на вопрос, *на какой исходной информации будут основаны наши выводы*. Исходную информацию целесообразно подразделять на: а) априорные сведения об искомых классах; б) предварительную выборочную информацию (так называемые обучающие выборки); в) наблюдения, подлежащие классификации. Определение необходимого для решения анализируемой задачи классификации математического инструментария (моделей, методов) в зависимости от характера и состава исходной информации может быть осуществлено с помощью табл. 12.3.

3. Если исследователь (статистик, эконометр) располагает, наряду с классифицируемыми данными, так называемыми *обучающими выборками* (12.1), то для решения задачи классификации он должен обратиться к методам **дискриминантного анализа** (см. п. 12.2). При этом каждый класс интерпретируется как одномодальная генеральная совокупность, закон распределения вероятностей (з.р.в.) которой оценивается по соответствующей обучающей выборке. Если априорные сведения позволяют сделать вывод об *общем параметрическом виде з.р.в.* каждого класса, то используют методы *параметрического дискриминантного анализа* (ДА). Если общий вид закона распределения внутри классов неизвестен, то обучающие выборки используются для получения непараметрических оценок внутриклассовых з.р.в., а сами процедуры классификации называются *непараметрическим ДА*.

5. Общая постановка *задачи классификации объектов в условиях отсутствия обучающих выборок* (т.е. *задачи автоматической классификации*) состоит в требовании разбиения этой совокупности на некоторое число (заранее известное или нет) однородных в определенном смысле классов. При этом априорные сведения об искомых классах могут давать основания для того, чтобы интерпретировать каждый искомый класс как параметрически заданную одномодальную генеральную совокупность, и тогда для построения правила классификации может быть использована *модель смеси распределений*. При более скудных априорных сведениях о классах используются методы и модели *кластер-анализа*, включая *иерархические методы классификации*. В обоих случаях *исходная статистическая информация* (наблюдения, подлежащие классификации) пред-

ставлены матрицами вида (9.1') «объект-свойство». В задачах кластер-анализа исходная статистическая информация может быть представлена также матрицей попарных расстояний (или близостей) вида (9.2). Понятие «однородности» основано на предположении, что геометрическая близость двух или нескольких объектов означает близость их «физических» состояний, их сходство.

6. Решить задачу расщепления смеси распределений (в выборочном варианте) — это значит по имеющейся выборке классифицируемых наблюдений, извлеченной из генеральной совокупности, являющейся смесью генеральных одномодальных совокупностей известного параметрического вида, построить статистические оценки для числа компонентов смеси, их удельных весов и параметров, их определяющих. После этого отнесение классифицируемых наблюдений к тому или иному классу производится по тем же правилам, что и в схеме дискриминантного анализа. *В теоретическом варианте* задача расщепления смеси заключается в восстановлении компонентов смеси и смешивающей функции (удельных весов) по заданному распределению всей (т. е. смешанной) генеральной совокупности и называется *задачей идентификации компонентов смеси* (эта задача не всегда имеет решение).

7. Базовая идея, лежащая в основе принятия решения, к какой из k анализируемых генеральных совокупностей следует отнести классифицируемое наблюдение, является одной и той же как для модели дискриминантного анализа (классификация при наличии обучения; см. п. 12.2), так и для модели смеси: и в том и в другом случае *наблюдение приписывают к той генеральной совокупности (к тому компоненту смеси), в рамках которой (которого) оно выглядит наиболее правдоподобным*. Однако главное отличие схемы параметрического ДА от схемы автоматической классификации, построенной на модели смеси распределений, — в способе оценивания неизвестных параметров, от которых зависят функции, описывающие классы (в первом случае — по обучающим выборкам, а во втором неизмеримо сложнее — в рамках одного из методов оценки параметров смеси распределений).

8. Основными «узкими местами» подхода, основанного на методе максимального правдоподобия статистического оценивания параметров смеси распределений, являются (помимо необходимости «угадать» общий параметрический вид распределения, задающего каждый из классов) требование ограниченности анализируемой функции правдоподобия, высокая сложность и трудоемкость процесса вычислительной реализации соответствующих процедур и медленная сходимость порождаемых ими итерационных алгоритмов.

9. *Математическая постановка задачи автоматической классификации* в рамках неиерархических схем кластер-анализа требует формализации понятия «качество разбиения». С этой целью в рассмотрение вводится понятие критерия (функционала) качества разбиения $Q(S)$, который задает способ сопоставления с каждым возможным разбиением S заданного множества объектов на классы некоторого числа $Q(S)$, оценивающего (в определенной шкале) степень оптимальности данного разбиения. Тогда задача поиска наилучшего разбиения S^* сводится к решению оптимизационной задачи вида

$$Q(S) \rightarrow \text{extr}, \\ S \in A,$$

где A — множество всех допустимых разбиений.

В статистической практике выбор функционала качества разбиения $Q(S)$ обычно осуществляется весьма произвольно и опирается скорее на эмпирические и профессионально-интуитивные соображения, чем на какую-либо точную формализованную схему.

10. В статистических классификационных *процедурах иерархического типа* главной целью анализа является получение наглядного представления о стратификационной структуре всей классифицируемой совокупности в *виде дендрограммы*.

11. *Выбор метрики (или меры близости) между объектами*, каждый из которых представлен значениями характеризующего его многомерного признака, является *узловым моментом исследования*, от которого решающим образом зависит окончательный вариант разбиения объектов на классы при любом используемом для этого алгоритме разбиения. В каждой конкретной задаче этот выбор должен производиться по-своему, в зависимости от главных целей исследования, физической и статистической природы анализируемого многомерного признака, априорных сведений о его вероятностной природе и т. д. В этом смысле схемы, основанные на анализе смесей распределений, а также классификация по исходным данным, уже представленным в виде матрицы попарных расстояний (близостей), находятся в выгодном положении, поскольку не требуют решения вопроса о выборе метрики.

12. Важное место в построении классификационных процедур, в первую очередь иерархических, занимает *проблема выбора способа вычисления расстояния между подмножествами объектов*. Полезное обобщение большинства используемых в статистической практике вариантов вычисления расстояний между двумя группами объектов дает расстояние, подсчитываемое как обобщенное степенное среднее всевозможных попарных расстояний между представителями рассматриваемых двух групп (см. (12.21)).

ГЛАВА 13. СНИЖЕНИЕ РАЗМЕРНОСТИ ИССЛЕДУЕМОГО МНОГОМЕРНОГО ПРИЗНАКА И ОТБОР НАИБОЛЕЕ ИНФОРМАТИВНЫХ ПОКАЗАТЕЛЕЙ

13.1. Сущность, типологизация и прикладная направленность задач снижения размерности

В исследовательской и практической статистической работе приходится сталкиваться с ситуациями, когда общее число p признаков $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, регистрируемых на каждом из множества обследуемых объектов (стран, городов, предприятий, семей, пациентов, технических или экологических систем), очень велико — порядка ста и более. Тем не менее имеющиеся многомерные наблюдения

$$X_i = \begin{pmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \vdots \\ x_i^{(p)} \end{pmatrix}, \quad i = 1, 2, \dots, n, \quad (13.1)$$

следует подвергнуть статистической обработке, осмыслить либо ввести в базу данных для того, чтобы иметь возможность их использовать в нужный момент.

Желание статистика представить каждое из наблюдений (13.1) в виде вектора Z некоторых *вспомогательных* показателей $z^{(1)}, z^{(2)}, \dots, z^{(p')}$ с существенно меньшим (чем p) числом компонент p' бывает обусловлено в первую очередь следующими причинами:

- необходимостью *наглядного представления* (визуализации) исходных данных (13.1), что достигается их проецированием на специально

подобранное трехмерное пространство ($p' = 3$), плоскость ($p' = 2$) или числовую прямую;

- стремлением к *лаконизму исследуемых моделей*, обусловленному необходимостью упрощения счета и интерпретации полученных статистических выводов;

- необходимостью существенного *сжатия объемов хранимой статистической информации* (без видимых потерь в ее информативности), если речь идет о записи и хранении массивов типа (13.1) в специальной базе данных.

При этом новые (вспомогательные) признаки $z^{(1)}, z^{(2)}, \dots, z^{(p')}$ могут выбираться из числа исходных или определяться по какому-либо правилу по совокупности исходных признаков, например как их линейные комбинации. При формировании новой системы признаков к последним предъявляются разного рода требования, такие, как наибольшая информативность (в определенном смысле), взаимная некоррелированность, наименьшее искажение геометрической структуры множества исходных данных и т. п. В зависимости от варианта формальной конкретизации этих требований приходим к тому или иному алгоритму снижения размерности. Имеется, по крайней мере, три основных типа принципиальных предположений, обуславливающих возможность перехода от большого числа p исходных показателей состояния (поведения, эффективности функционирования) анализируемой системы к существенно меньшему числу p' наиболее информативных переменных. Это, во-первых, *дублирование информации, доставляемой сильно взаимосвязанными признаками*; во-вторых, *неинформативность признаков, мало меняющихся при переходе от одного объекта к другому* (малая «вариабельность» признаков); в-третьих, *возможность агрегирования*, т. е. простого или «взвешенного» суммирования, по некоторым признакам.

Формально задача перехода (с наименьшими потерями в информативности) к новому набору признаков $\tilde{z}^{(1)}, \tilde{z}^{(2)}, \dots, \tilde{z}^{(p')}$ может быть описана следующим образом. Пусть $Z = Z(X) = (z^{(1)}, z^{(2)}, \dots, z^{(p')})^T$ — некоторая p' -мерная вектор-функция от исходных переменных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ ($p' \ll p$) и пусть $I_{p'}(Z(X))$ — определенным образом заданная мера информативности p' -мерной системы признаков $Z(X) = (z^{(1)}(X), \dots, z^{(p')}(X))^T$. Конкретный выбор функционала $I_{p'}(Z)$ зависит от специфики решаемой реальной задачи и опирается на один из возможных критериев: критерий *автоинформативности*, нацеленный на максимальное сохранение информации, содержащейся в исходном массиве $\{X_i\}_{i=1, \dots, n}$ относительно самих исходных признаков; и критерий *внешней информативности*, нацеленный

на максимальное «выжимание» из $\{X_i\}_{i=\overline{1,n}}$ информации, содержащейся в этом массиве относительно некоторых других (внешних) показателей.

Задача заключается в определении такого набора признаков $\tilde{Z}$, найденного в классе $F(X)$ допустимых преобразований исходных показателей $x^{(1)}, \dots, x^{(p)}$, что

$$I_{p'}(\tilde{Z}(X)) = \underset{Z \in F(X)}{\text{extr}} \left\{ I_{p'}(Z(X)) \right\}^1. \quad (13.2)$$

Тот или иной вариант конкретизации этой постановки (определяющий конкретный выбор меры информативности $I_{p'}(Z)$ и класса допустимых преобразований) приводит к конкретному методу снижения размерности: к методу главных компонент, факторному анализу, экстремальной группировке параметров и т. д.

Математическая модель, лежащая в основе построения того или иного метода снижения размерности, включает в себя обычно три основных компонента.

1. Форма задания исходной информации. Речь идет об ответе на следующие вопросы: а) в каком виде (т. е. в виде (9.1), (9.2) или еще каком-либо) задана описательная информация об объектах? б) имеется ли среди исходных статистических данных *обучающая* информация, т. е. какие-либо сведения об анализируемом результирующем свойстве? в) если обучающая информация присутствует в исходных статистических данных, то в какой именно форме она представлена? Это могут быть, в частности, в привязке к объекту O_i ($i = 1, 2, \dots, n$): значения «зависимой» количественной переменной («отклика») y_i в моделях регрессии; номер однородного по анализируемому свойству класса, к которому относится объект O_i в задаче классификации; порядковый номер (ранг) объекта O_i в ряду всех объектов, упорядоченных по степени проявления рассматриваемого свойства, в задачах анализа предпочтений и построения упорядоченных типологизаций; наконец, значения $Y_i = (y_i^{(1)}, \dots, y_i^{(q)})^T$ набора результирующих признаков, характеризующих анализируемое в классификационной задаче свойство².

¹ Речь идет о нахождении (в виде функции от X) такого вектора $Z(X) = (z^{(1)}(X), \dots, z^{(p')} (X))^T$, который обращает в максимум или минимум (в зависимости от конкретного содержательного смысла оптимизируемого критерия информативности) значение $I_{p'}(Z)$. Поэтому справа в данном соотношении записано extr («экстремум»).

² Перечисленные варианты не исчерпывают всех возможных форм представления обучающей информации. Так, например, при надлежащей интерпретации элементов

2. Тип оптимизируемого критерия $I_{p'}(Z)$ информативности искомого набора признаков $Z = (z^{(1)}, \dots, z^{(p')})^T$. Как уже отмечалось, критерий информативности может быть ориентирован на достижение разных целей.

Следует выделить целый класс *критериев автоинформативности*, т.е. критериев, оптимизация которых приводит к набору вспомогательных переменных $Z = (z^{(1)}, \dots, z^{(p')})^T$, позволяющих максимально точно воспроизводить (в том или ином смысле, в зависимости от конкретного вида критерия) информацию, содержащуюся в описательном массиве данных типа (9.1) или (9.2). Если описательная информация представлена в виде матрицы «объект – свойство» (9.1), то речь идет о максимально точном восстановлении $p \times n$ значений исходных переменных $x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)}$ по значениям существенно меньшего числа ($p' \times n$) вспомогательных переменных $z_i^{(1)}, \dots, z_i^{(p')}$. Если же описательная информация представлена в виде матрицы попарных сравнений объектов (9.2), то речь идет о максимально точном воспроизведении n^2 элементов этой матрицы (γ_{ij}) ($i, j = 1, \dots, n$) по значениям существенно меньшего числа ($p' \times n$) вспомогательных переменных $z_i^{(1)}, \dots, z_i^{(p')}$ ($i = 1, \dots, n$).

Будем называть *критериями внешней информативности* (имеется в виду информативность, внешняя по отношению к информации, содержащейся в описательном массиве (9.1) или (9.2)) такие критерии $I_{p'}(Z)$, которые нацелены на поиск экономных наборов вспомогательных переменных $Z(X) = (z^{(1)}(X), \dots, z^{(p')}(X))^T$, обеспечивающих максимально точное воспроизведение (по значениям Z , а значит в конечном счете по значениям X) информации, относящейся к результирующему признаку (варианты ее задания перечислены выше, в п. 1).

3. Класс $F(X)$ допустимых преобразований исходных признаков X . Вспомогательные признаки $Z = (z^{(1)}, \dots, z^{(p')})^T$ в случае представления исходной описательной информации в форме матрицы «объект – свойство» (т.е. в виде (9.1)) конструируются в виде функций от X , т.е. $Z = Z(X)$. Как обычно в таких ситуациях, чтобы обеспечить содержательность и конструктивную реализуемость решения оптимизационной задачи (13.2), следует предварительно договориться об ограниченном классе допустимых решений $F(X)$, в рамках которого эта оптимизационная задача будет решаться. Очевидно, от выбора $F(X)$ будет су-

γ_{ij} матрицы (9.2) она может быть также отнесена к разновидностям обучающей информации (если, скажем, γ_{ij} понимать как результат сравнения по анализируемому результирующему свойству объектов O_i и O_j).

щественно зависеть и получаемое решение $Z(X) = (z^{(1)}(X), \dots, z^{(p')}(X))^T$ упомянутой оптимизационной задачи.

Итак, следуя предложенной выше логике, мы должны были бы произвести типологизацию задач снижения размерности по трем «входам» (или «срезам»): форме задания исходной информации, типу (смыслу) оптимизируемого критерия информативности и классу допустимых преобразований исходных переменных. Однако в предлагаемой ниже форме представления результатов типологизации задач снижения размерности (табл. 13.1) эти принципы реализованы в упрощенном виде за счет следующих двух практических соображений: 1) подавляющее большинство методов снижения размерности базируется на *линейных* моделях, т. е. класс

Таблица 13.1

(АИ — автоинформативность, ВИ — внешняя информативность)

№ п/п	Класс и смысловая нацеленность критерия информативности; форма задания исходной информации	Название соответствующих моделей и методов. Главы и пункты книги
1	2	3
1	АИ: максимизация содержащейся в $z^{(1)}, \dots, z^{(p')}$ доли суммарной вариабельности исходных признаков $x^{(1)}, \dots, x^{(p)}$. Описательная информация: в форме (9.1). Обучающая информация: нет	Метод главных компонент, п. 13.2
2	АИ: Максимизация точности воспроизведения корреляционных связей между исходными признаками по их аппроксимациям с помощью вспомогательных переменных $z^{(1)}, \dots, z^{(p')}$. Описательная информация: в форме (9.1). Обучающая информация: нет	Модели и методы факторного анализа и главных компонент, п. 13.2 и 13.3
3	АИ: разбиение исходных признаков на группы высокоррелированных (внутри группы) переменных и отбор от каждой группы фактора, имеющего максимальную интегральную характеристику корреляционных связей со всеми признаками данной группы. Описательная информация: в форме (9.1). Обучающая информация: нет	Метод экстремальной группировки параметров, п. 13.4

Продолжение таблицы 13.1

1	2	3
4	АИ: приписывание каждому объекту O_i значений условных координат $(z_i^{(1)}, \dots, z_i^{(p')})$ таким образом, чтобы по ним максимально точно восстанавливалась заданная структура попарных описательных отношений между объектами. Описательная информация: в форме (9.2). Обучающая информация: нет	Многомерное шкалирование, п. 13.6
5	АИ: максимальное сохранение заданных описательным массивом (9.1) анализируемых структурно-геометрических и вероятностных свойств после его проецирования в пространство меньшей размерности (в пространство, натянутое на $z^{(1)}, \dots, z^{(p')}, p' < p$). Описательная информация: в форме (9.1). Обучающая информация: нет	Методы целенаправленного проецирования и отбор типобразующих признаков в кластер-анализе, метод главных компонент, п. 13.2 и гл. 12
6	ВИ: минимизация ошибки прогноза (восстановления) значения результирующей количественной переменной по значениям описательных переменных (предикторов). Описательная информация: в форме (9.1). Обучающая информация: в форме зарегистрированных на объектах $O_1, \dots, O_n$ значений соответственно $y_1, \dots, y_n$ результирующего количественного показателя y	Отбор существенных предикторов в регрессионном анализе, гл. 2 в томе 2
7	ВИ: минимизация вероятностей ошибочного отнесения объекта к одному из заданных классов по значениям его описательных переменных. Описательная информация: в форме (9.1). Обучающая информация: для каждого описанного с помощью (9.1) объекта указан номер класса, к которому он относится	Отбор типобразующих признаков в дискриминантном анализе, гл. 12

Продолжение таблицы 13.1

1	2	3
8	ВИ: максимизация точности воспроизведения (по значениям вспомогательных признаков) заданных в «обучении» отношений объектов по анализируемому результирующему свойству. Описательная информация: в форме (9.1). Обучающая информация: в форме попарных сравнений или упорядочений объектов по анализируемому результирующему свойству (см. сноску к п. 1 о возможности использования формы (9.2) для представления обучающей информации)	Методы латентно-структурного анализа, в том числе построение некой латентной характеристики изучаемого результирующего свойства, п. 13.5
9	ВИ: максимизация точности воспроизведения (по значениям условных вспомогательных переменных) заданных в «обучающей информации» попарных отношений объектов по анализируемому результирующему свойству. Описательная информация: нет. Обучающая информация: в форме (9.2) (см. сноску к п. 1)	Многомерное шкалирование как средство латентно-структурного анализа, п. 13.6

допустимых преобразований $F(X)$ — это класс линейных (как правило, подходящим образом нормированных) преобразований исходных признаков $x^{(1)}, \dots, x^{(p)}$; 2) спецификация формы задания исходной информации связана со спецификацией смысловой нацеленности критерия информативности, а поэтому их удобнее давать в общей графе.

Данная в табл. 13.1 типологизация, как и всякая иная классификация, не претендует на исчерпывающую полноту. Заметим, что пункт 9 этой таблицы повторяет по существу пункт 4, они отличаются только интерпретацией исходных данных вида (9.2) и соответственно конечными прикладными целями исследования.

13.2. Метод главных компонент

13.2.1. Основные понятия и определения

Во многих задачах обработки многомерных наблюдений и, в частности, в задачах классификации исследователя интересуют в первую оче-

редь лишь те признаки, которые обнаруживают наибольшую изменчивость (наибольший разброс) при переходе от одного объекта к другому.

С другой стороны, не обязательно для описания состояния объекта использовать какие-то из исходных, непосредственно замеренных на нем признаков. Так, например, для определения специфики фигуры человека при покупке одежды достаточно назвать значения двух признаков (размер-рост), являющихся производными от измерений ряда параметров фигуры. При этом, конечно, теряется какая-то доля информации (портной измеряет до одиннадцати параметров на клиенте), как бы огрубляются (при агрегировании) получающиеся при этом классы. Однако, как показали исследования, к вполне удовлетворительной классификации людей с точки зрения специфики их фигуры приводит система, использующая три признака, каждый из которых является некоторой комбинацией от большого числа непосредственно замеряемых на объекте параметров.

Именно эти принципиальные установки заложены в сущность того линейного преобразования исходной системы признаков, которое приводит к главным компонентам. Формализуются же эти установки следующим образом.

Следуя общей оптимизационной постановке задачи снижения размерности (13.2) и полагая анализируемый признак X p -мерной случайной величиной с вектором средних значений $a = (a^{(1)}, \dots, a^{(p)})$ и ковариационной матрицей $\Sigma = (\sigma_{ij})$ ($i, j = 1, 2, \dots, p$), вообще говоря, неизвестными, определим в качестве класса $F(X)$ допустимых преобразований исследуемых признаков $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ их всевозможные *линейные ортогональные нормированные комбинации*, т. е.

$$F = \left\{ Z: z^{(j)} = \sum_{\nu=1}^p c_{j\nu} (x^{(\nu)} - a^{(\nu)}), j = 1, 2, \dots, p \right\},$$

где

$$\sum_{\nu=1}^p c_{j\nu}^2 = 1 \quad \text{и} \quad \sum_{\nu=1}^p c_{j\nu} c_{k\nu} = 0 \quad (13.3)$$

для $j = 1, 2, \dots, p$ и $k = 1, 2, \dots, p$, но $j \neq k$, а в качестве критерия (меры) информативности p' -мерной системы показателей $Z(X) = (z^{(1)}(X), z^{(2)}(X), \dots, z^{(p')}(X))$ выражение

$$I_{p'}(Z(X)) = \frac{D z^{(1)} + \dots + D z^{(p')}}{D x^{(1)} + \dots + D x^{(p)}}. \quad (13.4)$$

Тогда при любом фиксированном $p' = 1, 2, \dots, p$ вектор искомых вспомогательных переменных $\tilde{Z}(X) = (\tilde{z}^{(1)}(X), \dots, \tilde{z}^{(p')}(X))^T$ определяется как такая линейная комбинация

$$\tilde{Z} = LX \quad (13.5)$$

(где

$$L = \begin{pmatrix} l_{11} & \dots & l_{1p} \\ \dots & \dots & \dots \\ l_{p'1} & \dots & l_{p'p} \end{pmatrix} -$$

матрица, строки которой удовлетворяют условию ортогональности), что

$$I_{p'}(\tilde{z}^{(1)}(X), \dots, \tilde{z}^{(p')}(X)) = \max_{Z(X) \in F} I_{p'}(Z(X)).$$

Полученные таким образом переменные $\tilde{z}^{(1)}(X), \dots, \tilde{z}^{(p)}(X)$ и называют главными компонентами вектора X . Отсюда вытекает следующее определение главных компонент.

Первой главной компонентой $\tilde{z}^{(1)}(X)$ исследуемой системы показателей $X = (x^{(1)}, \dots, x^{(p)})^T$ называется такая нормированно-центрированная линейная комбинация этих показателей, которая среди всех прочих нормированно-центрированных линейных комбинаций переменных $x^{(1)}, \dots, x^{(p)}$ обладает наибольшей дисперсией.

k -й главной компонентой $\tilde{z}^{(k)}(X)$ ($k = 2, 3, \dots, p$) исследуемой системы показателей $X = (x^{(1)}, \dots, x^{(p)})^T$ называется такая нормированно-центрированная линейная комбинация этих показателей, которая не коррелирована с $k-1$ предыдущими главными компонентами и среди всех прочих нормированно-центрированных и некоррелированных с предыдущими $k-1$ главными компонентами линейных комбинаций переменных $x^{(1)}, \dots, x^{(p)}$ обладает наибольшей дисперсией.

З а м е ч а н и е 1 (переход к центрированным переменным). Поскольку, как увидим ниже, решение задачи (а именно вид матрицы линейного преобразования L) зависит только от элементов ковариационной матрицы Σ , которые в свою очередь не изменяются при замене исходных переменных $x^{(j)}$ переменными $x^{(j)} - c^{(j)}$ ($c^{(j)}$ — произвольные постоянные числа), то в дальнейшем будем считать, что исходная система показателей уже *центрирована*, т.е. что $E x^{(j)} = 0$, $j = 1, 2, \dots, p$. В статистической практике этого добиваются, переходя к наблюдениям $\tilde{x}_i^{(j)} = x_i^{(j)} - \bar{x}^{(j)}$, где $\bar{x}^{(j)} = \sum_{i=1}^n x_i^{(j)} / n$ (для упрощения обозначений волнистую черту над цен-

трированной переменной и над главной компонентой в дальнейшем ставить не будем).

З а м е ч а н и е 2 (переход к выборочному варианту). Поскольку в реальных статистических задачах располагаем *лишь оценками* $\hat{\mathbf{a}}$ и $\hat{\Sigma}$ соответственно вектора средних $\mathbf{a}$ и ковариационной матрицы Σ , то во всех дальнейших рассуждениях под $\hat{\mathbf{a}}^{(j)}$ понимается $\bar{x}^{(j)}$, а под $\hat{\sigma}_{kj}$ — выборочная ковариация $\hat{\sigma}_{kj} = \sum_{i=1}^n (x_i^{(k)} - \bar{x}^{(k)})(x_i^{(j)} - \bar{x}^{(j)})/n$ ($j, k = 1, 2, \dots, p$).

З а м е ч а н и е 3. Использование главных компонент оказывается наиболее естественным и плодотворным в ситуациях, в которых все компоненты $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ исследуемого вектора X имеют *общую физическую природу* и соответственно *измерены в одних и тех же единицах*. К таким примерам можно отнести исследование структуры бюджета времени индивидуумов (все $x^{(i)}$ измеряются в единицах времени), исследование структуры потребления семей (все $x^{(i)}$ измеряются в денежных единицах), исследование общего развития и умственных способностей индивидуумов с помощью специальных тестов (все $x^{(i)}$ измеряются в баллах), разного рода антропологические исследования (все $x^{(i)}$ измеряются в единицах меры длины) и т. д. Если же признаки $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ измеряются в различных единицах, то результаты исследования с помощью главных компонент будут существенно зависеть от выбора масштаба и природы единиц измерения. Поэтому в подобных ситуациях исследователь предварительно переходит к вспомогательным безразмерным признакам $x^{*(i)}$, например, с помощью нормирующего преобразования

$$x_{\nu}^{*(i)} = \frac{x_{\nu}^{(i)}}{\sqrt{\hat{\sigma}_{ii}}} \quad \left(\begin{array}{l} i = 1, 2, \dots, p \\ \nu = 1, 2, \dots, n \end{array} \right),$$

где $\hat{\sigma}_{ii}$ соответствует ранее введенным обозначениям, а затем строит главные компоненты относительно этих вспомогательных признаков X^* и их ковариационной матрицы $\hat{\Sigma}_{X^*}$, которая, как легко видеть, является одновременно выборочной *корреляционной* матрицей $\hat{R}$ исходных наблюдений X_i .

13.2.2. Вычисление главных компонент

Из определения главных компонент следует, что для вычисления первой главной компоненты необходимо решить оптимизационную задачу ви-

да (13.2), т. е. в данном случае:

$$\begin{cases} D(l_1 X) \rightarrow \max_{l_1}; \\ l_1 l_1^T = 1, \end{cases} \quad (13.6)$$

где l_1 — первая строка матрицы L (см. (13.5)). Учитывая центрированность переменной X (т. е. $EX = 0$) и то, что $E(XX^T) = \Sigma$, имеем

$$D(l_1 X) = E(l_1 X)^2 = E(l_1 X X^T l_1^T) = l_1 \Sigma l_1^T.$$

Следовательно, задача (13.6) может быть записана

$$\begin{cases} l_1 \Sigma l_1^T \rightarrow \max_{l_1}; \\ l_1 l_1^T = 1, \end{cases} \quad (13.6')$$

Вводя функцию Лагранжа $\varphi(l_1, \lambda) = l_1 \Sigma l_1^T - \lambda(l_1 l_1^T - 1)$ и дифференцируя ее по компонентам вектор-столбца l_1^T , имеем (см. Приложение 2):

$$\frac{\partial \varphi}{\partial l_1^T} = 2 \Sigma l_1^T - 2 \lambda l_1^T,$$

что дает систему уравнений для определения l_1 :

$$(\Sigma - \lambda I) l_1^T = 0 \quad (13.7)$$

(здесь $0 = (0, 0, \dots, 0)^T$ — p -мерный вектор-столбец из нулей).

Для того чтобы существовало ненулевое решение системы (13.7) (а оно должно быть ненулевым, так как $l_1 l_1^T = 1$), матрица $\Sigma - \lambda I$ должна быть вырожденной, т. е.

$$|\Sigma - \lambda I| = 0. \quad (13.8)$$

Этого добиваются подбором соответствующего значения λ . Уравнение (13.8) (относительно λ) называется *характеристическим* для матрицы Σ . Известно (см. Приложение 2), что при симметричности и неотрицательной определенности матрицы Σ (каковой она и является как всякая ковариационная матрица) это уравнение имеет p вещественных неотрицательных корней $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, называемых *характеристическими* (или *собственными*) *значениями* матрицы Σ .

Учитывая, что $D\tilde{z}^{(1)} = D(l_1 X) = l_1 \Sigma l_1^T$ и $l_1 \Sigma l_1^T = \lambda$ (последнее соотношение следует из соотношения (13.7) после его умножения слева на l_1 , с учетом $l_1 l_1^T = 1$), получаем $D\tilde{z}^{(1)}(X) = \lambda$.

Поэтому для обеспечения максимальной величины дисперсии переменной $\tilde{z}^{(1)}$ нужно выбрать из p собственных значений матрицы Σ *наибольшее*, т. е.

$$D \tilde{z}^{(1)}(X) = \lambda_1.$$

Подставляем λ_1 в систему уравнений (13.7) и, решая ее относительно $l_{11}, \dots, l_{1p}$, определяем компоненты вектора l_1 .

Таким образом, *первая главная компонента получается как линейная комбинация $\tilde{z}^{(1)}(X) = l_1 X$, где l_1 — собственный вектор матрицы Σ , соответствующий наибольшему собственному числу этой матрицы.*

Далее аналогично можно показать, что $\tilde{z}^{(k)}(X) = l_k X$, где l_k — собственный вектор матрицы Σ , соответствующий k -му по величине собственному значению λ_k этой матрицы.

Таким образом, соотношения для определения всех p главных компонент вектора X могут быть представлены в виде (13.5), где $\tilde{Z} = (\tilde{z}^{(1)}, \dots, \tilde{z}^{(p)})^T$, $X = (x^{(1)}, \dots, x^{(p)})^T$, а матрица L состоит из строк $l_j = (l_{j1}, \dots, l_{jp})$, $j = \overline{1, p}$, являющихся собственными векторами матрицы Σ , соответствующими собственным числам λ_j . При этом сама матрица L в соответствии с условиями (13.3) является *ортогональной*, т. е.

$$LL^T = L^T L = I. \quad (13.9)$$

В дальнейшем в целях упрощения обозначений мы будем опускать «тильду» над переменными главных компонент, т. е. обозначать главные компоненты просто $Z = (z^{(1)}, z^{(2)}, \dots, z^{(p')})$.

13.2.3. Основные числовые характеристики главных компонент

Определим основные числовые характеристики (средние значения, дисперсии, ковариации) главных компонент в терминах основных числовых характеристик исходных переменных и собственных значений матрицы Σ :

а) $E Z = E(L X) = L \cdot E X = 0$;

б) ковариационная матрица вектора главных компонент:

$$\begin{aligned} \Sigma_Z &= E(ZZ^T) = E((L X)(L X)^T) = E(L X X^T L^T) \\ &= L \cdot E(X X^T) \cdot L^T = L \cdot \Sigma \cdot L^T. \end{aligned}$$

Умножая слева соотношения

$$(\Sigma - \lambda_k I) l_k^T = 0, \quad k = \overline{1, p},$$

на l_j ($j = \overline{1, p}$), получаем, что

$$\mathbf{L}\mathbf{\Sigma}\mathbf{L}^T = \mathbf{\Sigma}_Z = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p \end{pmatrix}. \quad (13.10)$$

Из (13.10), в частности, следует подтверждение взаимной некоррелированности главных компонент, а также $\mathbf{D} z^{(k)} = \lambda_k$ ($k = \overline{1, p}$);

в) сумма дисперсий исходных признаков равна сумме дисперсий всех главных компонент. Действительно,

$$\begin{aligned} \sum_{k=1}^p \mathbf{D} z^{(k)} &= \text{tr } \mathbf{\Sigma}_Z = \text{tr } (\mathbf{L}\mathbf{\Sigma}\mathbf{L}^T) = \text{tr } ((\mathbf{L}\mathbf{\Sigma})\mathbf{L}^T) \\ &= \text{tr } (\mathbf{L}^T(\mathbf{L}\mathbf{\Sigma})) = \text{tr } ((\mathbf{L}^T\mathbf{L})\mathbf{\Sigma}) = \text{tr } \mathbf{\Sigma} = \sum_{k=1}^p \mathbf{D} x^{(k)}; \end{aligned}$$

г) обобщенная дисперсия исходных признаков X ($\det \mathbf{\Sigma}_X$) равна обобщенной дисперсии главных компонент. Действительно, обобщенная дисперсия вектора Z равна

$$\begin{aligned} \det \mathbf{\Sigma}_Z &= \det (\mathbf{L}\mathbf{\Sigma}\mathbf{L}^T) = \det ((\mathbf{L}\mathbf{\Sigma})\mathbf{L}^T) = \det (\mathbf{L}^T(\mathbf{L}\mathbf{\Sigma})) \\ &= \det ((\mathbf{L}\mathbf{L}^T)\mathbf{\Sigma}) = \det (\mathbf{\Sigma}). \end{aligned}$$

Следствие 1. Из б) и в), в частности, следует, что критерий информативности метода главных компонент (13.4) может быть представлен в виде

$$I_{p'}(Z(X)) = \frac{\lambda_1 + \dots + \lambda_{p'}}{\lambda_1 + \dots + \lambda_p}, \quad (13.4')$$

где $\lambda_1, \lambda_2, \dots, \lambda_p$ — собственные числа ковариационной матрицы $\mathbf{\Sigma}$ вектора X , расположенные в порядке убывания.

Кстати, представление $I_{p'}(Z(X))$ в виде (13.4') дает исследователю некоторую основу, опорную точку зрения, при вынесении решения о том, сколько последних главных компонент можно без особого ущерба изъять из рассмотрения, сократив тем самым размерность исследуемого пространства.

Действительно, анализируя с помощью (13.4') изменение относительно доли дисперсии, вносимой первыми p' главными компонентами, в зависимости от числа этих компонент, можно разумно определить число

компонент, которое целесообразно оставить в рассмотрении. Так, при изменении $I_{p'}$, изображенном на рис. 13.1, очевидно, целесообразно было бы сократить размерность пространства с $p = 10$ до $p' = 3$, так как добавление всех остальных семи главных компонент может повысить суммарную характеристику рассеяния не более чем на 10%.

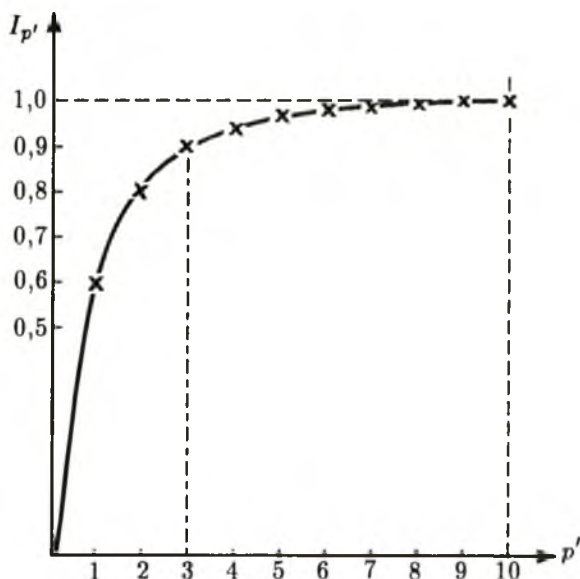


Рис. 13.1. Изменение относительной доли суммарной дисперсии исследуемых признаков, обусловленной первыми p' главными компонентами, в зависимости от p' (случай $p = 10$)

Следствие 2. Если X^* — вектор нормированных признаков $x^{*(1)}, \dots, x^{*(p)}$, т. е. $E x^{*(j)} = 0$ и $D x^{*(j)} = 1$ для $j = \overline{1, p}$, то согласно замечанию 3 ковариационная и корреляционные матрицы совпадают (т. е. $\Sigma_{X^*} = R$) и из б) и в) следует

$$\text{tr } \Sigma_Z = \text{tr } \Sigma_{X^*} = \text{tr } R = p$$

или

$$\sum_{k=1}^p \lambda_k = p.$$

Тогда критерий информативности (13.4') может быть представлен в виде

$$I_{p'}(Z(X)) = \frac{\lambda_1 + \dots + \lambda_{p'}}{p}. \quad (13.4'')$$

Проиллюстрируем вычисление и анализ главных компонент на примере.

Пример 13.1. При формировании типобразующих признаков предприятий отрасли были обследованы 24 предприятия ($n = 24$) по трем технико-экономическим показателям: объему выпускаемой продукции $x^{(1)}$, основным фондам $x^{(2)}$ и фонду оплаты труда $x^{(3)}$ (все переменные измерялись в денежных единицах). По полученным в результате обследования исходным статистическим данным $(x_i^{(1)}, x_i^{(2)}, x_i^{(3)})$, $i = 1, 2, \dots, 24$, была определена выборочная ковариационная матрица

$$\hat{\Sigma} = \begin{pmatrix} 451,39 & 271,17 & 168,70 \\ 271,17 & 171,73 & 103,29 \\ 168,70 & 103,29 & 66,65 \end{pmatrix}.$$

Решая, в соответствии с (13.8), кубическое уравнение (относительно λ) вида

$$\begin{vmatrix} 451,39 - \lambda & 271,17 & 168,70 \\ 271,17 & 171,73 - \lambda & 103,29 \\ 168,70 & 103,29 & 66,65 - \lambda \end{vmatrix} = 0,$$

находим $\lambda_1 = 680,40$, $\lambda_2 = 6,50$, $\lambda_3 = 2,86$.

Подставляя последовательно численные значения λ_1 , λ_2 и λ_3 в систему (13.7) и решая эти системы относительно неизвестных $l_i = (l_{i1}, l_{i2}, l_{i3})$ ($i = 1, 2, 3$), получаем

$$\begin{aligned} l_1 &= (0,8126 \quad 0,4955 \quad 0,3068), \\ l_2 &= (-0,5454 \quad 0,8321 \quad 0,1006), \\ l_3 &= (-0,2054 \quad -0,2491 \quad 0,9465). \end{aligned}$$

В качестве главных компонент получаем

$$\begin{aligned} z^{(1)} &= 0,81x^{(1)} + 0,50x^{(2)} + 0,31x^{(3)}, \\ z^{(2)} &= -0,55x^{(1)} + 0,83x^{(2)} + 0,10x^{(3)}, \\ z^{(3)} &= -0,21x^{(1)} - 0,25x^{(2)} + 0,95x^{(3)}. \end{aligned}$$

Здесь под $x^{(1)}$, $x^{(2)}$ и $x^{(3)}$ подразумеваются отклонения объема выпускаемой продукции ($x^{(1)}$), основных фондов ($x^{(2)}$) и фонда оплаты труда ($x^{(3)}$) предприятия от соответствующих средних значений.

Вычисление относительной доли суммарной дисперсии, обусловленной одной, двумя и тремя главными компонентами, в соответствии с фор-

мулой (13.4') дает

$$\begin{aligned} q(1) &= \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = 0,9864; \\ q(2) &= \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = 0,9958; \\ q(3) &= 1. \end{aligned}$$

Отсюда можно сделать вывод, что почти вся информация о специфике предприятия данного типа содержится в одной лишь первой главной компоненте, которую и естественно использовать при соответствующей типологизации предприятий.

Матрица «нагрузок» $A = (a_{ij})$, $i, j = 1, 2, \dots, p$, главных компонент на исходные признаки также является важной характеристикой главных компонент. Если анализируемые переменные $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^T$ предварительно процентрированы и пронормированы (см. выше замечания 1 и 3), т.е. если главные компоненты строятся для признаков $X^* = (x^{*(1)}, x^{*(2)}, \dots, x^{*(p)})^T$, $E x^{*(i)} = 0$, $D x^{*(i)} = 1$, $i = 1, 2, \dots, p$, то элементы матрицы нагрузок a_{ij} определяют одновременно степень тесноты парной линейной связи (т.е. парный коэффициент корреляции) между $x^{*(i)}$ и $z^{(j)}$ и удельный вес влияния *пронормированной* j -й главной компоненты на признак $x^{*(i)}$.

Матрица нагрузок A определяется соотношением

$$A = L^T \Lambda^{\frac{1}{2}}, \quad (13.11)$$

где

$$\Lambda^{\frac{1}{2}} = \Sigma_Z^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda_1} & & & 0 \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ 0 & & & \sqrt{\lambda_p} \end{pmatrix}. \quad (13.12)$$

Докажем, что элементы a_{ij} действительно обладают сформулированными выше свойствами. Введем в рассмотрение *нормированные* главные компоненты $Z_n = (z_n^{(1)}, z_n^{(2)}, \dots, z_n^{(p)})^T$, получающиеся из обычных главных компонент домножением на $\Lambda^{-\frac{1}{2}}$, т.е.

$$Z_n = \Lambda^{-\frac{1}{2}} Z. \quad (13.13)$$

Очевидно, что $z_n^{(j)} = z^{(j)} / \sqrt{\lambda_j}$ и соответственно

$$D z_n^{(j)} = 1. \quad (13.14)$$

Выразим из соотношения (13.5) X^* через Z :

$$X^* = L^{-1} Z = L^T Z \quad (13.15)$$

(при этом учтено, что мы, во-первых, упростили обозначения и вместо $\tilde{Z}$ теперь используем Z , а во-вторых, что матрица L — ортогональна по условию (см. (13.3)) и, следовательно, $L^{-1} = L^T$ (см. Приложение 2)).

Принимая во внимание (13.11) и (13.13), из (13.15) получаем:

$$X^* = L^T \Lambda^{\frac{1}{2}} \Lambda^{-\frac{1}{2}} Z = A Z_H, \quad (13.16)$$

т. е.

$$x^{*(i)} = a_{i1} z_H^{(1)} + a_{i2} z_H^{(2)} + \dots + a_{ip} z_H^{(p)}, \quad (13.16')$$

т. е. коэффициент a_{ij} действительно определяет удельный вес влияния j -й нормированной главной компоненты $z_H^{(j)}$ на i -й исходный признак.

Далее,

$$\begin{aligned} r(x^{*(i)}, z^{(j)}) &= \frac{E[(x^{*(i)} - E x^{*(i)})(z^{(j)} - E z^{(j)})]}{\sqrt{D x^{*(i)} D z^{(j)}}} \\ &= \frac{E(x^{*(i)} z^{(j)})}{\sqrt{\lambda_j}} = E\left(x^{*(i)} \frac{z^{(j)}}{\sqrt{\lambda_j}}\right) \\ &= E(x^{*(i)} z_H^{(j)}) = E\left[(a_{i1} z_H^{(1)} + \dots + a_{ip} z_H^{(p)}) z_H^{(j)}\right] \\ &= a_{ij} E\left((z_H^{(j)})^2\right) = a_{ij}, \end{aligned} \quad (13.17)$$

т. е. коэффициент a_{ij} действительно определяет величину парного коэффициента корреляции между $x^{*(i)}$ и $z^{(j)}$ (при выводе соотношения (13.17) использовались нормированность и центрированность переменных $x^{*(i)}$ и $z_H^{(j)}$, а также тот факт, что $z_H^{(j)} = z^{(j)} / \sqrt{\lambda_j}$).

Отметим еще два свойства элементов матрицы нагрузок A .

Из определения матрицы A (13.11) следует:

$$A^T A = (\Lambda^{\frac{1}{2}} L)(L^T \Lambda^{\frac{1}{2}}) = \Lambda^{\frac{1}{2}} (\Lambda L^T) \Lambda^{\frac{1}{2}} = \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} = \Lambda,$$

а это означает, что сумма квадратов элементов любого j -го столбца матрицы A равна дисперсии (j -й) главной компоненты λ_j , т. е.

$$a_{1j}^2 + a_{2j}^2 + \dots + a_{pj}^2 = \lambda_j. \quad (13.18)$$

Возводя в квадрат обе части соотношения (13.16') и беря математическое ожидание от результата, непосредственно имеем:

$$D x^{*(i)} = E (a_{i1} z_n^{(1)} + \dots + a_{ip} z_n^{(p)})^2 = \sum_{j=1}^p a_{ij}^2 E (z_n^{(j)})^2 = \sum_{j=1}^p a_{ij}^2, \quad (13.19)$$

т.е. с учетом $D x^{*(i)} = 1$ получаем, что *сумма квадратов элементов любой (i-й) строки матрицы нагрузок A равна единице.*

Приведенные здесь свойства используются, в частности, при содержательной интерпретации главных компонент. Так, соотношение (13.16') дает основание придавать главной компоненте $z^{(j)}$ содержательный смысл, соответствующий исходному признаку $x^{*(i_0)}$, для которого коэффициент a_{i_0j} достигает максимального значения (при условии, что $|a_{i_0j}| > 0,6$).

Пример 13.2. Компонентный анализ проведен по данным двадцати сельскохозяйственных районов ($n = 20$) области, которые содержат результаты измерений следующих показателей: $x^{(1)}$ — число колесных тракторов на 100 га; $x^{(2)}$ — число зерноуборочных комбайнов на 100 га; $x^{(3)}$ — число орудий поверхностной обработки почвы на 100 га; $x^{(4)}$ — количество удобрений, расходуемых на гектар; $x^{(5)}$ — количество средств защиты растений, расходуемых на гектар.

Требовалось выделить $p' < 5$ первых главных компонент для анализа и дать им содержательную интерпретацию.

Расчеты проводились по *нормированным данным* и представлены в табл. 13.2.

Таблица 13.2

Главные компоненты $z^{(i)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$	$z^{(4)}$	$z^{(5)}$
Собственные значения λ_i	3,04	1,41	0,43	0,10	0,02
Вклад i-й главной компоненты (%) в суммарную дисперсию	60,8	28,2	8,6	2,0	0,4
Суммарный вклад первых главных компонент (%)	60,8	89,0	97,6	99,6	100,0

При расчете относительного вклада главных компонент учитывалось, что $\sum_{i=1}^p \lambda_i = p = 5$. Для анализа были оставлены две первые главные компоненты ($p' = 2$), на которые приходится 89% суммарной вариации.

Для интерпретации главных компонент построена матрица факторных нагрузок

$$A = \begin{pmatrix} 0,95^* & 0,97^* & 0,94^* & 0,24 & 0,56 \\ -0,19 & -0,17 & -0,28 & 0,88^* & 0,67^* \end{pmatrix}^T.$$

Звездочкой (*) отмечены элементы $|a_{ij}| > 0,6$, которые следует учитывать при интерпретации главных компонент $z^{(1)}$ и $z^{(2)}$.

Из вида матрицы нагрузок **A** следует, что первая главная компонента наиболее тесно связана с показателями: $x^{(1)}$ — число колесных тракторов ($a_{11} = r(x^{(1)}, z^{(1)}) = 0,95$); $x^{(2)}$ — число зерноуборочных комбайнов ($a_{21} = r(x^{(2)}, z^{(1)}) = 0,97$); $x^{(3)}$ — число орудий поверхностной обработки почвы на 100 га ($a_{31} = r(x^{(3)}, z^{(1)}) = 0,94$). Поэтому первая главная компонента $z^{(1)}$ интерпретирована как уровень механизации работ.

Вторая главная компонента $z^{(2)}$ тесно связана с количествами удобрения ($x^{(4)}$) и средств защиты растений ($x^{(5)}$), расходуемых на гектар. Соответственно $z^{(2)}$ интерпретируется как уровень химизации растениеводства.

13.2.4. Геометрическая интерпретация главных компонент

Всякий переход к меньшему числу (p') переменных $z^{(1)}, \dots, z^{(p')}$, осуществляемый с помощью ортогонального линейного преобразования (матрицы) $\mathbf{C} = (c_{ij})$, $i = 1, 2, \dots, p'$, $j = 1, 2, \dots, p$, можно рассматривать как проекцию исследуемых p -мерных наблюдений $X_1, X_2, \dots, X_n$ в пространство размерности p' , натянутое на оси $Oz^{(1)}, Oz^{(2)}, \dots, Oz^{(p')}$, где

$$z^{(i)} = \sum_{j=1}^p c_{ij} x^{(j)}, \quad i = 1, 2, \dots, p'. \quad (13.20)$$

При этом проекциями p -мерных исходных наблюдений X_i ($i = 1, 2, \dots, n$) будут p' -мерные точки

$$Z_i = \mathbf{C} X_i, \quad i = 1, 2, \dots, n. \quad (13.21)$$

Для пояснения сущности того линейного преобразования исходной системы признаков, которое приводит к главным компонентам, рассмотрим его геометрическую интерпретацию на примере двумерной системы наблюдений $(x_i^{(1)}, x_i^{(2)})$, $i = 1, 2, \dots, n$, извлеченной из нормальной генеральной совокупности со средним значением $\mathbf{a} = (a^{(1)}, a^{(2)})$ и ковариационной матрицей

$$\Sigma = \begin{pmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \quad |r| \leq 1, \quad \sigma_1 > 0, \quad \sigma_2 > 0.$$

Здесь σ_1^2 и σ_2^2 — дисперсии компонент соответственно $x^{(1)}$ и $x^{(2)}$, а r — коэффициент корреляции между ними. Геометрически это означает, что

точки $(x_i^{(1)}, x_i^{(2)})$ будут располагаться примерно в очертаниях эллипсоидов рассеивания вида (см. рис. 13.2а)

$$\frac{1}{1-r^2} \left[\left(\frac{x^{(1)} - a^{(1)}}{\sigma_1} \right)^2 - 2r \left(\frac{x^{(1)} - a^{(1)}}{\sigma} \right) \times \left(\frac{x^{(2)} - a^{(2)}}{\sigma_2} \right) + \left(\frac{x^{(2)} - a^{(2)}}{\sigma_2} \right)^2 \right] = c^2.$$

В этом случае для изучения $(x^{(1)}, x^{(2)})$ удобно перейти к новым координатам $(z^{(1)}, z^{(2)})$ с помощью преобразования:

$$\begin{aligned} z^{(1)} &= (x^{(1)} - a^{(1)}) \cos \alpha + (x^{(2)} - a^{(2)}) \sin \alpha, \\ z^{(2)} &= -(x^{(1)} - a^{(1)}) \sin \alpha + (x^{(2)} - a^{(2)}) \cos \alpha, \end{aligned}$$

где

$$\operatorname{tg} 2\alpha = \frac{2r\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}.$$

После этого преобразования точки $(z_i^{(1)}, z_i^{(2)})$ также будут распределены нормально, но компонента $z^{(1)}$ уже не будет зависеть от $z^{(2)}$. Кроме того, если выбрать направления так, что $D z^{(1)} \geq D z^{(2)}$, то геометрически это будет означать следующее: сначала производится перенос начала координат в точку $(a^{(1)}, a^{(2)})$, а затем оси поворачиваются на угол α так, чтобы ось $z^{(1)}$ шла вдоль главной оси эллипсоида рассеивания (рис. 13.2а). Чем ближе $|r|$ к единице, тем теснее группируются наблюдения около главной оси эллипсоида рассеивания (т.е. около новой оси $z^{(1)}$) и тем менее значащим для исследователя является разброс точек в направлении оси $z^{(2)}$, а следовательно, и сама эта координата. В предельном случае $|r| = 1$, исследуемые наблюдения в координатах $(z^{(1)}, z^{(2)})$ вообще не отличаются по координате $z^{(2)}$ (см. рис. 13.2б).

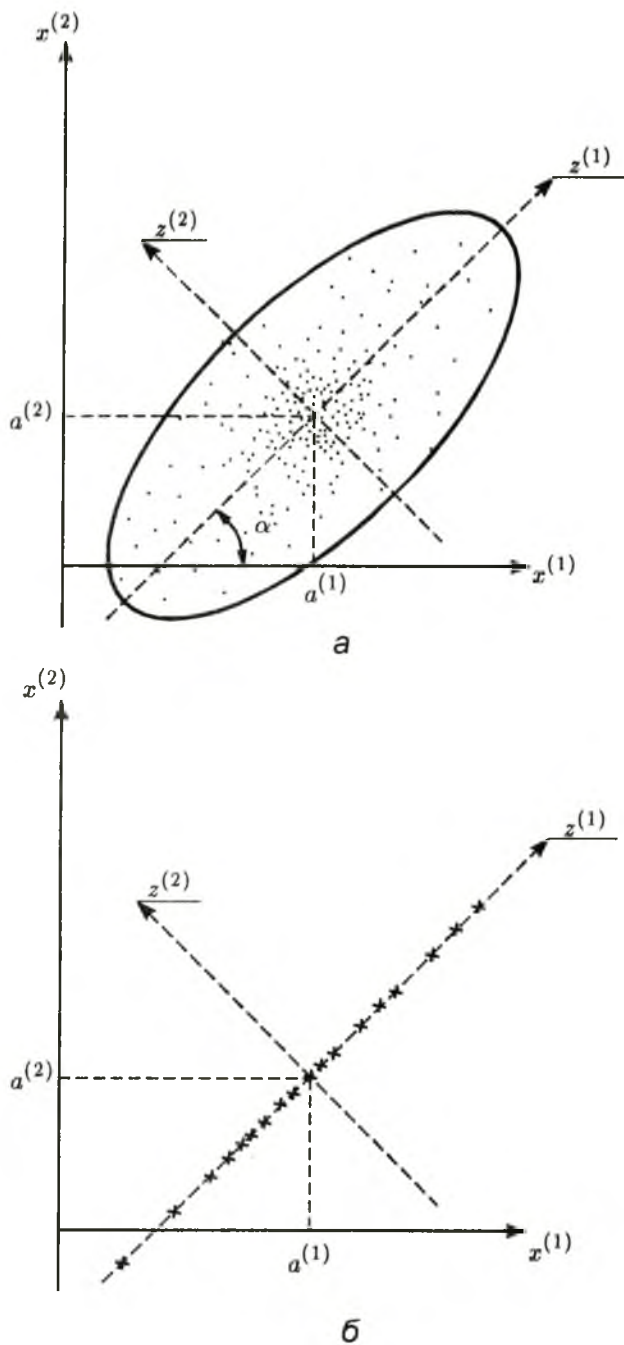


Рис. 13.2. Эллипс рассеяния исследуемых наблюдений и направление координатных осей главных компонент $z^{(1)}$ и $z^{(2)}$: а) умеренный разброс точек; б) отсутствие разброса точек в направлении второй главной компоненты (вырожденный случай)

13.2.5. Оптимальные свойства главных компонент

Описываемые ниже свойства первых p' главных компонент во многом объясняют их широкую распространенность в практике статистических, в том числе эконометрических, исследований. Оказывается, к главным компонентам можно прийти, решая оптимизационные задачи, на первый взгляд не имеющие ничего общего с оптимизационной задачей типа (13.2).

Свойство наименьшей ошибки «автопрогноза» или наилучшей самовоспроизводимости. Можно показать, что с помощью первых p' главных компонент $z^{(1)}, z^{(2)}, \dots, z^{(p')}$ ($p' < p$) исходных признаков $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ достигается наилучший прогноз этих признаков среди всех прогнозов, которые можно построить с помощью p' линейных комбинаций набора из p произвольных признаков.

Поясним и уточним сказанное. Пусть требуется заменить исходный исследуемый p -мерный вектор наблюдений X на вектор $Z = (z^{(1)}, z^{(2)}, \dots, z^{(p')})^T$ меньшей размерности p' , в котором каждая из компонент являлась бы линейной комбинацией p исходных (или каких-либо других, вспомогательных) признаков, теряя при этом не слишком много информации. Информативность нового вектора Z зависит от того, в какой степени p' введенных вспомогательных переменных дают возможность «реконструировать» p исходных (измеряемых на объектах) признаков с помощью подходящих линейных комбинаций $z^{(1)}, z^{(2)}, \dots, z^{(p')}$. Естественно полагать, что ошибка прогноза X по Z (обозначим ее σ) будет определяться так называемой остаточной дисперсионной матрицей вектора X при вычитании из него наилучшего прогноза по Z , т.е. матрицей $\Delta = (\Delta_{ij})$, где

$$\Delta_{ij} = \mathbf{E} \left\{ \left(x^{(i)} - \sum_{l=1}^{p'} b_{il} z^{(l)} \right) \left(x^{(j)} - \sum_{l=1}^{p'} b_{jl} z^{(l)} \right) \right\}.$$

Здесь $\sum_{l=1}^{p'} b_{il} z^{(l)}$ — наилучший, в смысле метода наименьших квадратов, прогноз $x^{(i)}$ по компонентам $z^{(1)}, z^{(2)}, \dots, z^{(p')}$ (см. гл. 2 тома 2). Ошибка прогноза X по Z задается как некоторая определенная функция от элементов матрицы Δ , т.е. $\sigma = f(\Delta)$, где $f(\Delta)$ определяет некоторый критерий качества предсказания.

Рассмотрим следующие естественные меры ошибки прогноза:

$$f(\Delta) = \text{tr}(\Delta) = \Delta_{11} + \Delta_{22} + \dots + \Delta_{pp}; \quad (13.22)$$

$$f(\Delta) = \|\Delta\| = \sqrt{\sum_{i=1}^p \sum_{j=1}^p \Delta_{ij}^2}. \quad (13.23)$$

Здесь $\text{tr}(\Delta)$ и $\|\Delta\|$ — соответственно *след* и *евклидова норма* матрицы Δ . Доказано, что функции (13.22) и (13.23) одновременно достигают минимума тогда и только тогда, когда в качестве $z^{(1)}, z^{(2)}, \dots, z^{(p')}$ выбраны первые p' главных компонент вектора X , причем величина ошибки прогноза σ явным образом выражается через последние $p - p'$ собственных чисел исходной ковариационной матрицы Σ или *приблизленно* — через последние $p - p'$ собственных чисел $\lambda_{p'+1}, \dots, \lambda_p$ выборочной ковариационной матрицы $\hat{\Sigma}$, построенной по наблюдениям $X_1, X_2, \dots, X_n$. В частности,

$$\text{при } f(\Delta) = \text{tr}(\Delta): \sigma \approx \lambda_{p'+1} + \lambda_{p'+2} + \dots + \lambda_p;$$

$$\text{при } f(\Delta) = \|\Delta\|: \sigma \approx \sqrt{\lambda_{p'+1}^2 + \lambda_{p'+2}^2 + \dots + \lambda_p^2}.$$

Поясним идею описания (прогноза) исходных признаков $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ с помощью меньшего, чем p , числа их линейных комбинаций на примере 13.1.

В этом примере $p = 3$. Зададимся целью снизить размерность исходного факторного пространства до единицы ($p' = 1$), т.е. описать все три признака с помощью линейных комбинаций только от одной вспомогательной переменной.

В соответствии с описанным выше экстремальным свойством «автопрогноза» главных компонент возьмем в качестве этой единственной вспомогательной переменной первую главную компоненту, т.е. переменную

$$z^{(1)} = 0,81x^{(1)} + 0,50x^{(2)} + 0,31x^{(3)}.$$

Метод наименьших квадратов приводит к следующему правилу вычисления неизвестных коэффициентов b_{i1} (см. гл. 2 тома 2):

$$\begin{aligned} b_{i1} &= \frac{\text{cov}(x^{(i)}, z^{(1)})}{D z^{(1)}} \\ &= \frac{0,81 \text{cov}(x^{(1)}, x^{(1)}) + 0,50 \text{cov}(x^{(2)}, x^{(1)}) + 0,31 \text{cov}(x^{(3)}, x^{(1)})}{D z^{(1)}}. \end{aligned}$$

Подставляя в эту формулу значения $\text{cov}(x^{(i)}, x^{(j)})$, взятые из ковари-

ационной матрицы Σ примера 13.1, получаем

$$\begin{aligned}x^{(1)} &= b_{11}z^{(1)} + \varepsilon^{(1)} = 0,805z^{(1)} + \varepsilon^{(1)}, \\x^{(2)} &= b_{21}z^{(1)} + \varepsilon^{(2)} = 0,49z^{(1)} + \varepsilon^{(2)}, \\x^{(3)} &= b_{31}z^{(1)} + \varepsilon^{(3)} = 0,310z^{(1)} + \varepsilon^{(3)},\end{aligned}$$

где $\varepsilon^{(i)}$ — случайные (остаточные) ошибки прогноза исходных центрированных компонент по первой главной компоненте $z^{(1)}$.

Если в качестве относительной ошибки прогноза исходного признака $x^{(i)}$ по первой главной компоненте $z^{(1)}$ рассмотреть величину $\delta_i = (D\varepsilon^{(i)} / D x^{(i)}) \cdot 100\%$, то несложные подсчеты дают $\delta_1 = 2\%$, $\delta_2 = 1,2\%$ и $\delta_3 = 0,8\%$.

Суммарная характеристика относительной ошибки прогноза признаков $x^{(1)}$, $x^{(2)}$ и $x^{(3)}$ по $z^{(1)}$ (в соответствии с вышеописанным) может быть подсчитана по формуле

$$\delta_{\text{сум.отн}} = 100\% \cdot \frac{\text{tr}(\Delta)}{D(x^{(1)} + x^{(2)} + x^{(3)})} = 100\% \cdot \frac{\lambda_2 + \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} = 1,36\%.$$

Свойства наименьшего искажения геометрической структуры множества исходных p -мерных наблюдений при их проектировании в пространство p' первых главных компонент. Речь идет о следующих трех оптимальных свойствах главных компонент (формулируются без доказательства).

Свойство 1. *Сумма квадратов расстояний от исходных точек-наблюдений $X_1, X_2, \dots, X_n$ до пространства, натянутого на первые p' главных компонент, наименьшая относительно всех других подпространств размерности p' , полученных с помощью произвольного линейного преобразования исходных координат.*

Наглядным пояснением этого свойства может служить рис. 13.2а, на котором ось $z^{(1)}$ соответствует подпространству, натянутому на первую главную компоненту (т.е. $p = 2$ и $p' = 1$), а сумма квадратов расстояний до этого подпространства есть сумма перпендикуляров, опущенных из точек, изображающих наблюдения $X_i = (x_i^{(1)}, x_i^{(2)})$, на эту ось.

Свойство 2. *Среди всех подпространств заданной размерности p' ($p' < p$), полученных из исследуемого признакового пространства с помощью произвольного линейного преобразования исходных координат $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, в подпространстве, натянутом на первые p' главных компонент, наименее искажается сумма квадратов расстояний между всевозможными парами рассматриваемых точек-наблюдений.*

В данном свойстве за критерий наименьшего искажения геометрической структуры совокупности исходных наблюдений $X_1, X_2, \dots, X_n$ принимается величина

$$M_p(X) = \sum_{i=1}^n \sum_{j=1}^n (X_i - X_j)^T (X_i - X_j),$$

т. е. сумма квадратов евклидовых расстояний между всевозможными парами имеющихся наблюдений.

После проектирования точек X_i в p' -мерное пространство, определяемое матрицей преобразования C , мы получим точки-проекции $Z_i = C X_i$ ($i = 1, 2, \dots, n$) и соответствующую им сумму квадратов евклидовых расстояний

$$M_{p'}(Z(C)) = \sum_{i=1}^n \sum_{j=1}^n (Z_i - Z_j)^T (Z_i - Z_j).$$

Можно показать, что при $p' < p$ $M_p(X) \geq M_{p'}(Z(C))$.

Так вот, на преобразовании L , с помощью которого получают первые p' главных компонент, достигается минимум разности $M_p(X) - M_{p'}(Z(C))$, т. е.

$$M_p(X) - M_{p'}(Z(L)) = \min_C [M_p(X) - M_{p'}(Z(C))].$$

Свойство 3. Среди всех подпространств заданной размерности p' ($p' < p$), полученных из исследуемого факторного пространства с помощью произвольного линейного преобразования исходных координат $x^{(1)}, \dots, x^{(p)}$, в пространстве, натянутом на первые p' главных компонент, наименее искажаются расстояния от рассматриваемых точек-наблюдений до их общего «центра тяжести», а также углы между прямыми, соединяющими всевозможные пары точек-наблюдений с их общим «центром тяжести».

13.2.6. Статистические свойства выборочных главных компонент; статистическая проверка некоторых гипотез

Смысл математико-статистических методов, как известно, состоит в том, чтобы по некоторой части исследуемой генеральной совокупности (по выборке или, что то же, по ограниченному ряду наблюдений $X_1, X_2, \dots, X_n$) выносить обоснованные суждения о ее свойствах в целом.

Применительно к рассматриваемой задаче нас в первую очередь интересует, как сильно свойства и характеристики *выборочных* главных

компонент могут отличаться от соответствующих свойств и характеристик главных компонент *всей генеральной совокупности* и, в частности, как эта мера отличия зависит от объема выборочной совокупности (n), по которой эти выборочные главные компоненты были построены. Так, например, для изучения природы внутренних связей между характеристиками различных статей семейного бюджета потребления и для выявления небольшого числа наиболее существенных в этом смысле показателей исследователь может обследовать какое-то количество (n) семей и по полученным результатам наблюдения $X_1, X_2, \dots, X_n$ построить главные компоненты $\hat{z}^{(1)}, \hat{z}^{(2)}, \dots, \hat{z}^{(p')}$. Однако, увеличивая объем выборки n , т. е. добавляя к имеющимся наблюдениям результаты наблюдения по дополнительно обследованным семьям, естественно ожидать, что пересчет главных компонент *с учетом добавленных наблюдений*, вообще говоря, изменит (хотя, быть может, и незначительно) ранее полученные значения интересующих нас характеристик: $\hat{\lambda}_i, \hat{l}_i$ ($i = 1, 2, \dots, p$) и т. п. В то же время существует, по-видимому, такое (столь большое) n , дальнейшее увеличение которого уже не будет практически приводить к изменению основных характеристик главных компонент (другими словами, мы вправе ожидать, что в соответствии со свойством статистической устойчивости (см. п. 6.2) главные компоненты выборок достаточно большого объема практически совпадают с главными компонентами всей генеральной совокупности).

Выяснению некоторых вопросов, связанных с оценкой близости различных выборочных $(\hat{z}^{(i)}, \hat{l}_i, \hat{\lambda}_i)$ и теоретических $(z^{(i)}, l_i, \lambda_i)$ характеристик главных компонент, и посвящен настоящий пункт. Приведенные ниже результаты исследований неизменно опираются на допущение нормальности исследуемой генеральной совокупности и взаимной независимости извлеченных из нее наблюдений. Как и прежде, под $X_1, X_2, \dots, X_n$ будем понимать *центрированные* наблюдения, которые, строго говоря, даже при независимых исходных наблюдениях уже не будут независимыми. Однако при достаточно больших n можно пренебречь этим эффектом нарушения независимости. Таким образом, $X_i \in N(0, \Sigma)$, $i = 1, 2, \dots, n$ (как следует из предыдущего, вектор средних значений $a = EX$ определяет лишь точку в p -мерном пространстве, в которую переносится начало координат при переходе к главным компонентам, и с самого начала будем считать этот перенос уже осуществленным).

Вспомогательные факты, относящиеся к свойствам выборочных характеристик главных компонент (см., например, [Андерсон]). Если все характеристические корни $\lambda_1, \lambda_2, \dots, \lambda_p$ ковариационной матрицы Σ различны, что и имеет место в большинстве приложений ана-

лиза главных компонент, то справедливо следующее:

1) *характеристические корни $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_p$ и соответствующие им собственные векторы $\hat{l}_1, \hat{l}_2, \dots, \hat{l}_p$ выборочной ковариационной матрицы $\hat{\Sigma}$ являются оценками максимального правдоподобия для соответствующих теоретических характеристик (соответственно $\lambda_1, \lambda_2, \dots, \lambda_p$ и $l_1, l_2, \dots, l_p$) и обладают всеми хорошими свойствами этих оценок (состоятельность, асимптотическая эффективность). Следовательно, выборочные главные компоненты $\hat{z}^{(i)} = \hat{l}_i X$ ($i = \overline{1, p}$) можно интерпретировать как оценки главных компонент $z^{(i)}$ всей генеральной совокупности. Если среди характеристических корней $\lambda_1, \lambda_2, \dots, \lambda_p$ встречаются равные между собой, то оценки максимального правдоподобия для λ_i и l_i определяются иначе. Аналогичные результаты имеют место и при оценке характеристических корней и соответствующих им собственных векторов корреляционной матрицы;*

2) *величины $\sqrt{n-1}(\hat{\lambda}_i - \lambda_i)$ ($i = \overline{1, p}$) асимптотически (по $n \rightarrow \infty$) нормальны со средним значением 0 и с дисперсией, равной $2\lambda_i^2$, и независимы от других выборочных характеристических корней;*

3) *выборочный характеристический корень $\hat{\lambda}_i$ распределен асимптотически (по $n \rightarrow \infty$) независимо от компонент соответствующего ему собственного вектора $\hat{l}_i$ ($i = \overline{1, p}$);*

4) *ковариация между r -й компонентой выборочного собственного вектора $\hat{l}_i$ и q -й компонентой выборочного собственного вектора $\hat{l}_j$ равна величине*

$$-\frac{\lambda_i \lambda_j l_{ir} l_{jq}}{(n-1)(\lambda_i - \lambda_j)^2}.$$

В заключение приведем два факта, относящихся к ситуациям, в которых компоненты нормального вектора наблюдений X взаимно независимы:

5) *пусть $X \in N(a, \Sigma)$, где ковариационная матрица имеет диагональный вид, т. е. $\text{cov}(x^{(i)}, x^{(j)}) = 0$ при $i \neq j$, $i, j = 1, 2, \dots, p$. И пусть $|\hat{R}| = \det(\hat{r}_{ij})$ — определитель выборочной корреляционной матрицы, построенной по наблюдениям $(X_1, \dots, X_n)$. Тогда при достаточно больших n ($n \rightarrow \infty$) статистика критерия отношения правдоподобия для проверки гипотезы о диагональном виде Σ может быть определена в виде $\gamma = -(n - \frac{2p+11}{6}) \ln |\hat{R}|$, а для ее функции распределения справедливо приближенное соотношение*

$$P\{\gamma \leq u\} \approx P\left\{\chi^2\left(\frac{p(p-1)}{2}\right) \leq u\right\}$$

при относительной ошибке, не превосходящей сотых долей процента;

6) пусть наблюдения X_j извлечены из так называемой сферической p -мерной нормальной совокупности $N(a; \sigma^2 I)$, т. е. компоненты каждого из векторов X_j взаимно независимы и имеют одинаковые дисперсии $D x_j^{(i)}$, равные σ^2 . Тогда ковариационная матрица $\Sigma = \sigma^2 I$ имеет единственный корень (кратности p), оценкой максимального правдоподобия которого является величина

$$\hat{\lambda} = \frac{1}{pn} \sum_{i=1}^p \sum_{j=1}^n (x_j^{(i)} - \bar{x}^{(i)})^2,$$

причем величина $\hat{\lambda}/\sigma^2$ распределена по закону $\chi^2(p(n-1))$.

Статистика критерия отношения правдоподобия для проверки гипотезы о сферичности распределения исследуемого вектора наблюдений имеет вид

$$\omega = \frac{|n\hat{\Sigma}|}{[\frac{1}{p} \text{tr}(n\hat{\Sigma})]^p}$$

и при достаточно больших n ($n \rightarrow \infty$)

$$P \left\{ - \left(n - 1 - \frac{2p^2 + p + 2}{6p} \right) \ln \omega < t \right\} \approx P \left\{ \chi^2 \left(\frac{p(p+1)}{2} - 1 \right) < t \right\}$$

при относительной ошибке данного приближенного соотношения, не превосходящей сотых долей процента.

13.2.7. Применение свойств выборочных характеристик главных компонент

Опишем некоторые методы построения разного рода интервальных оценок для интересующих нас неизвестных характеристик главных компонент и критерии статистической проверки гипотез, относящихся к этим характеристикам:

1) *интервальная оценка (доверительный интервал) для i -го характеристического корня λ_i* получается (при больших n) с учетом асимптотической нормальности статистики $\sqrt{n-1}(\hat{\lambda}_i - \lambda_i)$:

$$\frac{\hat{\lambda}_i}{1 + u_{1-\frac{\alpha}{2}} \sqrt{\frac{2}{n-1}}} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{1 - u_{1-\frac{\alpha}{2}} \sqrt{\frac{2}{n-1}}}, \quad (13.24)$$

где данное неравенство справедливо с вероятностью $1 - \alpha$ (величиной α заранее задаемся), а u_q — q -квантиль стандартного нормального распределения (находится из таблиц Приложения 1).

Возвращаясь к примеру 13.1, по формуле (13.24) находим 95%-ный ($\alpha = 0,05$) доверительный интервал для наименьшего характеристического корня λ_3 по его выборочному значению $\hat{\lambda}_3 = 2,86$. В этом случае $n = 24$, $u_{0,975} = 1,96$, так что $1,81 \leq \lambda_3 \leq 6,78$.

Возможно обобщение асимптотического (по $n \rightarrow \infty$) доверительного интервала на случай кратных, т.е. повторяющихся корней. Если r — кратность корня λ_i , то $100(1 - \alpha)\%$ -ный доверительный интервал для неизвестного значения λ_i задается неравенством

$$\frac{\bar{\lambda}_i}{1 + u_{\frac{\alpha}{2}} \sqrt{\frac{2}{(n-1)r}}} \leq \lambda_i \leq \frac{\bar{\lambda}_i}{1 - u_{\frac{\alpha}{2}} \sqrt{\frac{2}{(n-1)r}}}, \quad (13.25)$$

где

$$\bar{\lambda}_i = \frac{1}{r} (\hat{\lambda}_i + \hat{\lambda}_{i+1} + \dots + \hat{\lambda}_{i+r-1}).$$

Вопрос о том, что неизвестный характеристический корень λ_i имеет кратность i , в частности, кратность, равную r , может быть решен с помощью следующего критерия:

2) *проверка гипотезы о равенстве нескольких (а именно r) характеристических корней*: $\lambda_i = \lambda_{i+1} = \dots = \lambda_{i+r-1}$. Очевидно, альтернативой этой гипотезе является утверждение, что не все корни среди $\lambda_i, \lambda_{i+1}, \dots, \lambda_{i+r-1}$ равны между собой. Оказывается, в предположении справедливости проверяемой гипотезы статистика

$$\gamma_r = (n-1) \sum_{j=1}^{i+r-1} \ln \hat{\lambda}_j + (n-1)r \ln \left(\frac{1}{r} \sum_{j=i}^{i+r-1} \hat{\lambda}_j \right)$$

распределена (асимптотически по $n \rightarrow \infty$) по закону «хи-квадрат» с $r(r+1)/2 - 1$ степенью свободы. Поэтому гипотеза $\lambda_i = \lambda_{i+1} = \dots = \lambda_{i+r-1}$ отвергается (с вероятностью ошибиться, равной α), если

$$\gamma_r > \chi_{\alpha}^2 \left(\frac{r(r+1)}{2} - 1 \right),$$

где $\chi_{\alpha}^2(m)$ — $100\alpha\%$ -ная точка χ^2 -распределения с m степенями свободы.

Особый интерес может представить специальный случай $i = p - r + 1$, т.е. проверка гипотезы о равенстве *последних* r собственных значений λ ,

что будет означать независимость и сферичность r последних признаков исследуемого вектора наблюдений.

Возвратимся к примеру 13.1. Тот факт, что оценка второго собственного значения ($\lambda_2 = 6,50$) попадает в доверительный интервал для λ_3 (см. выше), приводит к мысли, что, возможно, $\lambda_2 = \lambda_3$. Проверим эту гипотезу. В данном случае $n = 24$, $p = 3$, $i = 2$, $r = 2$, так что

$$\gamma_2 = -23(\ln 6,50 + \ln 2,86) + 46 \ln \frac{6,50 + 2,86}{2} = 3,70.$$

А поскольку $\chi_{0,05}^2(2) = 5,99$ и, следовательно, $\gamma_2 < \chi_{0,05}^2(2)$, то гипотезу $\lambda_2 = \lambda_3$ следует принять. Но тогда следует пересчитать доверительный интервал для $\lambda_2 = \lambda_3$ с учетом его кратности (в соответствии с (13.25)). Несложные подсчеты (при $\alpha = 0,05$ и соответственно $u_{1-\frac{\alpha}{2}} = u_{0,975} = 1,96$) дают: $2,62 \leq \lambda_2 \leq 6,21$, последнее неравенство будет справедливо в среднем в 95 случаях из 100;

3) *проверка гипотезы о независимости признаков $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, являющихся компонентами вектора наблюдений X* . Такая проверка нужна для установления целесообразности применения метода главных компонент: если признаки являются взаимно независимыми, то переход к главным компонентам сведется, по существу, лишь к упорядочению исходных признаков по принципу убывания их дисперсий. Воспользуемся статистикой критерия отношения правдоподобия для проверки гипотезы о диагональном виде ковариационной матрицы с целью проверки независимости компонент вектора наблюдений в следующем примере.

Пример 13.3. Исследовалось время, затрачиваемое работниками швейной фабрики на выполнение различных элементов операции глаженья одежды. Эту операцию можно разделить на следующие шесть элементов:

- 1) одежда размещается на гладильной доске ($x^{(1)}$);
- 2) разглаживаются короткие швы ($x^{(2)}$);
- 3) одежда перекладывается на гладильной доске ($x^{(3)}$);
- 4) разглаживаются длинные швы на три четверти ($x^{(4)}$);
- 5) разглаживаются остатки длинных швов ($x^{(5)}$);
- 6) одежду вешают на вешалку ($x^{(6)}$).

В этом случае X_ν представляет собой вектор измерений элементов ν -й выполненной операции. Компонента $x^{(i)}$ — это время, затраченное на выполнение i -го элемента операции, $n = 76$. Данные (время в секундах) обработаны, получены выборочные вектор среднего значения $\hat{a}$ и корре-

ляционная матрица $\hat{\mathbf{R}}$:

$$\hat{\mathbf{a}} = \begin{pmatrix} 9,47 \\ 25,56 \\ 13,25 \\ 31,44 \\ 27,29 \\ 8,70 \end{pmatrix};$$

$$\hat{\mathbf{R}} = \begin{pmatrix} 1,000 & 0,088 & 0,334 & 0,191 & 0,173 & 0,123 \\ 0,088 & 1,000 & 0,186 & 0,383 & 0,262 & 0,040 \\ 0,334 & 0,186 & 1,000 & 0,343 & 0,144 & 0,080 \\ 0,191 & 0,384 & 0,343 & 1,000 & 0,375 & 0,142 \\ 0,173 & 0,262 & 0,144 & 0,375 & 1,000 & 0,334 \\ 0,123 & 0,040 & 0,080 & 0,142 & 0,334 & 1,000 \end{pmatrix}.$$

Для исследователей представляет интерес проверка гипотезы о взаимной независимости шести случайных величин. Часто при изучении затрат времени предлагается новая операция, в которой элементы комбинируются иным способом. В новой операции некоторые элементы могут повторяться по несколько раз, а некоторые могут быть выброшены. Если оказываются независимыми величины, обозначающие время, затрачиваемое на различные элементы операции, то естественно считать, что и в новой операции они останутся независимыми. Тогда распределение затрат времени на новую операцию можно будет оценить, пользуясь средними значениями и дисперсиями, вычисленными для остальных элементов. Кроме того, нас интересует возможность выделения небольшого количества вспомогательных признаков (двух-трех), с помощью которых можно производить некоторую содержательную классификацию исследуемых работников (в том или ином смысле).

В этой задаче статистика критерия отношения правдоподобия, определенная в соответствии с п. 5, имеет вид: $\gamma = -(n - \frac{2p+11}{6}) \ln |\hat{\mathbf{R}}| = -\frac{433}{6} \ln 0,472 = 54,1$, а $p(p-1)/2 = 15$. Задавшись уровнем значимости критерия $\alpha = 0,01$ (вероятность ошибочно отвергнуть проверяемую гипотезу), находим (из таблиц) величину 1%-ной точки χ^2 -распределения с 15 степенями свободы: $\chi^2_{0,01}(15) = 30,6$. Поскольку $\gamma > \chi^2_{0,01}(15)$, то гипотезу следует отвергнуть, т.е. приходим к выводу, что значения затрат времени на различные элементы операции нельзя считать независимыми.

* * *

Прикладная направленность компонентного анализа отчасти продемонстрирована приведенными в данном пункте примерами, но в основном

отражена в табл. 13.1. В частности, из оптимальных свойств главных компонент следует, что они оказываются полезным статистическим инструментарием в задачах «автопрогноза» большого числа анализируемых показателей по сравнительно малому числу вспомогательных (латентных) переменных, визуализации многомерных данных, построения типобразующих признаков; при типологизации многомерных объектов, при предварительном анализе геометрической и вероятностной природы массива исходных данных (это отражено в п. 1, 2, 5 упомянутой таблицы). Мы увидим в гл. 2 и 4 тома 2, что к методу главных компонент обращаются и при построении различного рода регрессионных моделей.

13.3. Факторный анализ

13.3.1. Сущность модели факторного анализа

Развиваемые в рамках модели факторного анализа методы исходят из общей базовой идеи, в соответствии с которой структура связей между p анализируемыми признаками $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ может быть объяснена тем, что все эти переменные зависят (линейно или как-то иначе) от меньшего числа других, непосредственно не измеряемых («скрытых», «латентных») факторов $f^{(1)}, f^{(2)}, \dots, f^{(p')}$ ($p' < p$), которые принято называть общими и которые в большинстве моделей конструируются так, чтобы они оказались взаимно некоррелированными. При этом в общем случае, естественно, не постулируется возможность однозначного (детерминированного) восстановления значений каждого из наблюдаемых признаков $x^{(j)}$ по соответствующим значениям общих факторов $f^{(1)}, \dots, f^{(p')}$ (в предположении, что мы их умеем вычислять): допускается, что каждый из исходных признаков $x^{(j)}$ зависит также и от некоторой своей («специфической» только для него) остаточной случайной компоненты $u^{(j)}$, которая и обуславливает *статистический* характер связи между $x^{(j)}$, с одной стороны, и $f^{(1)}, \dots, f^{(p')}$ — с другой.

Конечная цель статистического исследования, проводимого с привлечением аппарата факторного анализа, как правило, состоит в выявлении и интерпретации латентных общих факторов с одновременным противоречивым стремлением минимизировать как их число, так и степень зависимости $x^{(j)}$ от своих специфических остаточных случайных компонент $u^{(j)}$. Как и в любой модельной схеме, эта цель может быть достигнута лишь приближенно.

В некотором смысле искомые общие факторы $f^{(1)}, \dots, f^{(p')}$ можно считать *причинами*, а наблюдаемые признаки $x^{(1)}, \dots, x^{(p)}$ — *следствиями*. Принято считать статистический анализ такого рода успешным, если большое число следствий удалось объяснить малым числом причин.

Таким образом, методы и модели факторного анализа нацелены, так же как и метод главных компонент, на сжатие информации или, что то же, на снижение размерности исходного признакового пространства. При этом из трех упомянутых в п. 13.1 предпосылок возможности снижения размерности (взаимная коррелированность исходных признаков, малая «вариабельность» некоторых из них, агрегирование) методы факторного анализа базируются в основном на первой.

Возникновение схемы и моделей факторного анализа обязано задачам психологии, относится к началу двадцатого века и связано с именами Ч. Спирмэна, Л. Тэрстоуна, Г. Томсона¹. Однако в силу ряда исторических причин и, в частности, из-за субъективных пристрастий и специфических научных интересов первых исследователей, работавших в данной области, вероятностно-статистические аспекты этого раздела многомерного статистического анализа долгое время оставались практически неразработанными, а интерпретации и анализу различных моделей факторного анализа была присуща некоторая неопределенность. Лишь с середины 50-х годов начинают появляться интересные результаты именно вероятностно-статистических исследований этого аппарата, среди которых работа [Anderson T. W., Rubin H.] признается основополагающей.

Параграф посвящен описанию канонической линейной модели факторного анализа. Весьма обширный (и достаточно сложный) материал, касающийся *подробностей вычислительной реализации* разнообразных версий модели факторного анализа, обусловленных различными вариантами априорных допущений о ее структуре (см. ниже п. 13.3.5), остается при этом вне рамок нашего описания. Читателю, интересующемуся этими аспектами проблемы, можно порекомендовать обратиться, например, к [Харман Г.].

13.3.2. Общий вид линейной модели, ее связь с главными компонентами

Как и прежде, будем для удобства полагать исследуемые наблюдения $X_1, X_2, \dots, X_n$ центрированными. Переход от исходных наблюдений

¹ В качестве первой опубликованной по этой теме работы называют обычно статью Spearman C. General intelligence objectively determined and measured /Amer. J. Psychol, 1904, vol. 15, pp. 201–293.

$X_1, X_2, \dots, X_n$ к центрированным осуществляется с помощью простого переноса начала координат в «центр тяжести» исходного множества наблюдений. Тогда линейная версия модели факторного анализа представляется в виде соотношений

$$\left. \begin{aligned} X &= \mathbf{Q} F + U, \\ \text{или в покомпонентной записи} \\ x^{(i)} &= \sum_{j=1}^{p'} q_{ij} f^{(j)} + u^{(i)} \quad (i = \overline{1, p}) \end{aligned} \right\}. \quad (13.26)$$

Здесь $\mathbf{Q} = (q_{ij})$ — прямоугольная матрица размера $p \times p'$ коэффициентов линейного преобразования (нагрузок общих факторов на исследуемые признаки), связывающего исследуемые признаки $x^{(i)}$ с ненаблюдаемыми (скрытыми) общими факторами $f^{(1)}, \dots, f^{(p')}$, а вектор-столбец $U = (u^{(1)}, \dots, u^{(p')})^T$ определяет ту часть каждого из исследуемых признаков, которая не может быть объяснена общими факторами, в том числе $u^{(i)}$ включает в себя, как правило, ошибки измерения признака $x^{(i)}$ (сравнить с (13.16)).

Применительно к каждому конкретному наблюдению X_ν ($\nu = 1, 2, \dots, n$) соотношение (13.26) дает

$$\left. \begin{aligned} X_\nu &= \mathbf{Q} F_\nu + U_\nu, \\ \text{или в покомпонентной записи} \\ x_\nu^{(i)} &= \sum_{j=1}^{p'} q_{ij} f_\nu^{(j)} + u_\nu^{(i)} \quad (i = \overline{1, p}; \nu = \overline{1, n}) \end{aligned} \right\}. \quad (13.26')$$

Будем предполагать, что вектор остаточных специфических факторов U подчиняется p -мерному нормальному распределению $N(0, \mathbf{V})$, не зависит от F и состоит из взаимно независимых компонент, т.е. его ковариационная матрица $\mathbf{V} = \mathbf{E}(UU^T)$ имеет диагональный вид, где по диагонали стоят элементы $v_{ii} = \mathbf{D} u^{(i)}$.

Вектор общих факторов $F = (f^{(1)}, \dots, f^{(p')})^T$, в зависимости от содержания конкретной задачи, может интерпретироваться либо как p' -мерная нормальная случайная величина со средним $\mathbf{E} F = \mathbf{0}$ (в силу центрированности исходных наблюдений) и с ковариационной матрицей специального вида $\mathbf{E}(FF^T) = \mathbf{I}_{p'}$, либо как вектор неизвестных неслучайных параметров, вспомогательных переменных, значения которых меняются от наблюдения к наблюдению. При последней интерпретации вектора общих факторов более правильной является запись модели в виде (13.26'), причем условия центрированности, независимости и нормированности дисперсий

компонент вектора F в этом случае имеют вид:

$$\frac{1}{n} \sum_{\nu=1}^n F_{\nu} = \mathbf{0}, \quad \frac{1}{n} \sum_{\nu=1}^n F_{\nu} F_{\nu}^{\top} = \mathbf{I}_{p'}.$$

Однако при обоих вариантах интерпретации вектора общих факторов F исследуемый вектор наблюдений X оказывается нормально распределенной p -мерной случайной величиной: при первом варианте как линейная комбинация двух нормальных случайных векторов (F и U), а при втором варианте за счет нормальности специфических факторов $u^{(i)}$. При этом из (13.26) и из сделанных выше допущений немедленно следует, что

$$\left. \begin{aligned} E x^{(i)} &= 0 \\ \left\{ \begin{aligned} \sigma_{ii} &= \sum_{\nu=1}^{p'} q_{i\nu}^2 + v_{ii} \\ \sigma_{ij} &= \sum_{\nu=1}^{p'} q_{i\nu} q_{j\nu} \end{aligned} \right. & (i, j = \overline{1, p}), \end{aligned} \right\} \quad (13.27)$$

или в матричной записи

$$E X = \mathbf{0}, \quad \Sigma = \mathbf{Q} \mathbf{Q}^{\top} + \mathbf{V}$$

Примером достаточно прозрачной интерпретации модели факторного анализа может служить ее формулировка в терминах так называемых интеллектуальных тестов. При этом наблюдение по признаку $x_j^{(i)}$ выражает отклонение оценки, например, в баллах, данной j -му индивидууму на экзамене по i -му тесту, от некоторого среднего уровня. Естественно предположить, что в качестве ненаблюдаемых общих факторов $f^{(1)}, \dots, f^{(p')}$, от которых будут зависеть оценки индивидуумов по всем p тестам, выступят такие факторы, как характеристика *общей одаренности* индивидуума $f^{(1)}$, характеристики его *математических* $f^{(2)}$, *технических* $f^{(3)}$ или *гуманитарных* $f^{(4)}$ способностей.

Отметим, что соотношения (13.26) формально *воспроизводят* запись модели множественной регрессии (см. гл. 10), в которой под $f^{(i)}$ ($i = 1, 2, \dots, p'$) понимаются так называемые объясняющие переменные (факторы-аргументы). Однако принципиальное отличие модели факторного анализа от регрессионных схем состоит в том, что переменные $f^{(i)}$, выступающие в роли аргументов в моделях регрессии, *не являются непосредственно наблюдаемыми в моделях факторного анализа*, в то время как в регрессионном анализе значения $f^{(i)}$ измеряются на статистически обследованных объектах.

З а м е ч а н и е о связи метода главных компонент и метода факторного анализа. Рассмотрим следующую общую схему, включающую в себя в качестве частных случаев обе сравни-

главная компонента X ; при этом i -й столбец матрицы A_m ($i = 1, \dots, m$) есть $\sqrt{\lambda_i} l_i^T$, где l_i — собственный вектор матрицы Σ , соответствующий характеристическому корню λ_i .

Если же определение аппроксимирующего вектора $\hat{X}(m) = A_m F(m)$ подчинить идее максимального объяснения корреляции между исходными признаками $x^{(i)}$ и $x^{(j)}$ с помощью вспомогательных (ненаблюдаемых) факторов $f^{(1)}(m), f^{(2)}(m), \dots, f^{(m)}(m)$ и, в частности, идее минимизации величины

$$\sum_{\substack{i,j=1 \\ (i \neq j)}}^p \left[\frac{\text{cov}(x^{(i)}, x^{(j)}) - \text{cov}(\hat{x}^{(i)}(m), \hat{x}^{(j)}(m))}{(\sigma_{ii} \sigma_{jj})^{\frac{1}{2}}} \right]^2 \quad (13.29)$$

при условии неотрицательности величин $\sigma_{ii} - D \hat{x}^{(i)}$, то можно показать, что i -я строка оптимальной в этом смысле матрицы преобразования A_m состоит из m факторных нагрузок общих факторов $f^{(1)}(m), \dots, f^{(m)}(m)$ на i -й исходный признак $x^{(i)}$ в модели факторного анализа вида (13.26). Другими словами, сущность задачи минимизации (по A_m и $F(m)$) величины (13.29) состоит в следующем. Первый из m общих факторов $f^{(1)}(m)$ находится из условия, чтобы попарные корреляции между исходными признаками были как можно меньше, если влияние на них этого фактора $f^{(1)}(m)$ учтено. Следующий общий фактор $f^{(2)}(m)$ находится из условия максимального ослабления попарных корреляционных связей между исходными признаками, оставшихся после учета влияния первого общего фактора $f^{(1)}(m)$, и т. д.

Из сказанного, в частности, следует, что методы главных компонент и факторного анализа должны давать близкие результаты в тех случаях, когда главные компоненты строятся по корреляционным матрицам исходных признаков, а остаточные дисперсии v_{ii} сравнительно невелики.

13.3.3. Основные задачи факторного анализа

При разработке моделей факторного анализа приходится последовательно анализировать и решать следующие вопросы.

Существование модели. Далеко не для всякой заданной структуры связей между исходными признаками $X = (x^{(1)}, \dots, x^{(p)})^T$ можно (при заданном $p' < p$) построить модель факторного анализа, т. е. указать такие общие факторы $f^{(1)}, \dots, f^{(p')}$ (или доказать их существование), которые объясняли бы имеющуюся корреляцию между различными парами $x^{(i)}$ и $x^{(j)}$ в рамках модели (13.26). При каком характере связей между исход-

ными признаками $x^{(1)}, \dots, x^{(p)}$, т.е. при каких корреляционных (ковариационных) матрицах $\mathbf{R} = (r_{ij})$ ($\Sigma = (\sigma_{ij})$), а также при каком соотношении между числом наблюдаемых признаков p и числом скрытых общих факторов p' ($p' < p$) сделанное допущение о наличии определенных связей между $x^{(i)}$ ($i = 1, 2, \dots, p$), с одной стороны, и $f^{(j)}$ ($j = 1, 2, \dots, p'$) — с другой, является обоснованным и содержательным — в этом и заключается вопрос существования модели. В терминах модели (13.26) это означает, что не всякая ковариационная матрица Σ допускает представление вида (13.27), а следовательно, не всякий вектор наблюдений X допускает интерпретацию в рамках модели факторного анализа типа (13.26). Очевидно, условия представимости вектора наблюдений X в рамках модели факторного анализа должны формулироваться в терминах свойств ковариационной матрицы Σ , а также в виде некоторых соотношений между размерностью исходного пространства p и числом общих факторов p' .

Единственность (идентификация) модели. Оказывается, что если p , Σ и p' таковы, что допускают построение модели факторного анализа, то определение соответствующих факторов $F^T = (f^{(1)}, \dots, f^{(p')})$ и коэффициентов линейного преобразования $\mathbf{Q} = (q_{ij})$, связывающего X и F , не единственно. Спрашивается, при каких дополнительных ограничениях на матрицу преобразования $\mathbf{Q}$ и на ковариационную матрицу $\mathbf{V} = (v_{ij})$ остаточных специфических факторов $u^{(1)}, \dots, u^{(p)}$ определение параметров исходной модели факторного анализа будет единственным?

Алгоритмическое определение структурных параметров модели. При заданной ковариационной матрице Σ исходных признаков и известном числе общих факторов p' (и в предположении, что решение задачи определения структурных параметров $\mathbf{Q}$ и $\mathbf{V}$ существует) как конкретно вычислить неизвестные параметры модели?

Статистическое оценивание (по наблюдениям $X_1, X_2, \dots, X_n$ и при заданном p') *неизвестных структурных параметров модели.*

Статистическая проверка ряда гипотез, связанных с природой модели и значениями ее структурных параметров, таких, как гипотеза об истинном числе p' общих факторов, гипотеза адекватности принятой модели по отношению к имеющимся результатам наблюдения, гипотеза о значимом отличии от нуля интересующих нас коэффициентов q_{ij} линейного преобразования $\mathbf{Q}$ и т. п.

Построение статистических оценок для значений ненаблюдаемых общих факторов $f_i^{(1)}, f_i^{(2)}, \dots, f_i^{(p')}$ ($i = 1, 2, \dots, n$).

Далее мы кратко остановимся на решении некоторых из этих задач.

13.3.4. Вопросы идентификации модели факторного анализа

Будем в дальнейшем предполагать, что имеется по меньшей мере одно решение $(\mathbf{Q}, \mathbf{V})$ уравнений (13.27). При исследовании вопроса единственности решения системы (13.27) относительно $(\mathbf{Q}, \mathbf{V})$ (при заданных σ_{ij}) следует различать два аспекта проблемы. Во-первых, надо понять, при каких дополнительных условиях на искомую матрицу нагрузок $\mathbf{Q}$ и на соотношение между p и p' не может существовать двух различных решений $Q^{(1)}$ и $Q^{(2)}$, таких, чтобы одно из них нельзя было бы получить из другого с помощью соответствующим образом подобранного ортогонального преобразования $\mathbf{C}$ (единственность с точностью до ортогонального преобразования или с точностью до вращения факторов). Оказывается, достаточным условием единственности такого рода является требование к матрице $\mathbf{Q}$, чтобы при вычеркивании из нее любой строки оставшуюся матрицу можно было бы разделить на две подматрицы ранга p' , откуда автоматически следует требование $p' < (p - 1)/2$.

Можно показать, что для $p' = 1$ и $p' = 2$ это условие является одновременно и необходимым, откуда, в частности, следует, что случаи $p = 2$, $p' = 1$ и $p = 4$, $p' = 2$ не допускают идентификации модели факторного анализа в указанном выше смысле.

Будем предполагать далее, что имеется по меньшей мере одно решение $(\mathbf{Q}, \mathbf{V})$ системы (13.27) и что оно единственно с точностью до ортогонального преобразования.

Вставляя в уравнения (13.27) вместо найденного решения $(\mathbf{Q}, \mathbf{V})$ другую пару матриц $(\mathbf{Q}\mathbf{C}, \mathbf{V})$, где $\mathbf{C}$ — матрица (размера $p' \times p'$) любого ортогонального преобразования, легко убедиться, что и она (эта пара матриц) удовлетворяет данной системе уравнений. Следовательно, возвращаясь к модели (13.26), получаем, что наряду с общими факторами $F = (f^{(1)}, \dots, f^{(p')})^T$ можно рассмотреть (при тех же нагрузках q_{ij}) общие факторы $Z = \mathbf{C}F$. Поскольку, как известно (см. Приложение 2), ортогональное преобразование координат F геометрически означает вращение осей $f^{(1)}, \dots, f^{(p')}$ около начала координат на некоторый угол, то получается, что при отсутствии дополнительных условий на природу искомой матрицы нагрузок $\mathbf{Q}$ общие факторы $f^{(1)}, \dots, f^{(p')}$ могут быть определены лишь с точностью до вращения системы координат в соответствующем p' -мерном пространстве. Существует несколько вариантов дополнительных условий на класс матриц $\mathbf{Q}$, в котором следует искать решение системы (13.27), обеспечивающих уже окончательную однозначность решения $(\mathbf{Q}, \mathbf{V})$. От конкретного содержания этих условий зависит и способ численного выявления структуры искомой модели и соответственно способ

статистического оценивания неизвестных параметров q_{ij} , v_{ii} и факторов $f^{(i)}$. Поэтому остановимся на них параллельно с описанием методов статистического исследования модели факторного анализа.

13.3.5. Статистическое исследование модели факторного анализа

Итак, в распоряжении исследователя — последовательность многомерных наблюдений $X_1, X_2, \dots, X_n$, и с помощью модели (13.26) нужно перейти от исходных коррелированных признаков $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, являющихся компонентами каждого из наблюдений, к меньшему числу некоррелированных вспомогательных признаков (общих факторов) $f^{(1)}, \dots, f^{(p')}$. Для этого надо суметь определить оценки неизвестных нагрузок $\hat{q}_{ij}$ остаточных дисперсий $\hat{v}_{ii}$ и, наконец, самих общих факторов $\hat{f}^{(i)}$.

Как упоминалось, в основной модели (13.26) при $p' > 1$ оказывается слишком много неизвестных параметров для однозначного определения. Поэтому вначале исследователь должен выбрать какую-то систему дополнительных априорных соотношений, связывающих неизвестные параметры модели, которые делают решение задачи однозначным и позволяют получить относительно простое частное решение системы (13.27). Затем он может отказаться от этих дополнительных соотношений, подбирая с помощью подходящего ортогонального преобразования (вращения осей) тот вариант оценок нагрузок $\hat{q}_{ij}$ и остаточных дисперсий $\hat{v}_{ii}$, который ему кажется предпочтительнее в основном в отношении возможности *содержательной интерпретации* получаемых при этом общих факторов и их нагрузок. Остановимся подробнее на основных этапах статистического исследования модели факторного анализа.

Варианты дополнительных априорных соотношений между q_{ij} и v_{ii} , постулируемых исследователем с целью однозначной идентификации анализируемой модели:

- 1) решение (Q, V) системы (13.27) лежит лишь в классе таких матриц Q и V , для которых матрица $Q^T V Q$ имеет диагональный вид, причем диагональные элементы ее различны и упорядочены в порядке убывания;
- 2) из всех решений системы (13.27) выбирается лишь то, для которого матрица $Q^T Q$ диагональна, причем все диагональные элементы различны и упорядочены (в порядке убывания);
- 3) решение системы (13.27) ищут лишь среди таких матриц Q , которые для заранее заданной матрицы (размера $p \times p'$) $B = (b_{ij})$, $i = 1, \dots, p$,

$j = 1, \dots, p'$, ранга p' удовлетворяют требованию

$$\mathbf{B}^T \mathbf{Q} = \mathbf{D} = \begin{pmatrix} d_{11} & 0 & \dots & 0 \\ d_{21} & d_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ d_{p'1} & d_{p'2} & \dots & d_{p'p'} \end{pmatrix}.$$

В частности, выбор

$$\mathbf{B}^T = \begin{pmatrix} \overbrace{10\dots 0}^{p'} & \overbrace{0\dots 0}^{p-p'} \\ 01\dots 0 & 0\dots 0 \\ \dots & \dots \\ 00\dots 1 & 0\dots 0 \end{pmatrix}$$

приводит к ограничению на $\mathbf{Q}$ типа

$$\mathbf{Q} = \begin{pmatrix} q_{11} & 0 & 0 & \dots & 0 \\ q_{21} & q_{22} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ q_{p'1} & q_{p'2} & q_{p'3} & \dots & q_{p'p'} \\ \dots & \dots & \dots & \dots & \dots \\ q_{p1} & q_{p2} & q_{p3} & \dots & q_{pp'} \end{pmatrix},$$

что означает: первый исходный признак $x^{(1)}$ должен выражаться только через один первый общий фактор $f^{(1)}$, второй признак $x^{(2)}$ — через два общих фактора $f^{(1)}$ и $f^{(2)}$ и т.д.

Содержательную интерпретацию условий 3) следует искать в ситуациях, когда исследователь располагает некоторой априорной информацией, из которой можно, во-первых, извлечь реальный гипотетический смысл общих факторов и, во-вторых, постулировать наличие определенного числа нулевых элементов в матрице нагрузок $\mathbf{Q}$ (с более или менее точным указателем их «адреса»), что означает априорное отрицание зависимости исходных признаков $x^{(1)}$ от некоторых их общих факторов $f^{(j)}$ ($j = 1, 2, \dots, p'$). Эта же идея реализуется и в других, менее формализованных вариантах дополнительных условий («простые структуры», «нулевые элементы в специфических позициях»), на которых здесь не будем останавливаться.

Описание общего итерационного подхода к выявлению структуры модели факторного анализа. Конкретная реализация этого подхода зависит от выбора варианта идентифицирующих условий типа 1)–3).

Как правило, исследователю известна лишь оценка $\hat{\Sigma}$ ковариационной матрицы Σ . Поэтому в дальнейшем мы будем писать Σ , подразумевая, что это в действительности ее выборочное значение.

Логическая схема итераций следующая:

- задаемся некоторым нулевым приближением $V^{(0)}$ матрицы V ;
- используя (13.27), получаем нулевое приближение $\Psi^{(0)} = \Sigma - V^{(0)}$ матрицы $\Psi = QQ^T = \Sigma - V$;
- по Ψ с помощью некоторого специального приема (см. ниже) последовательно определяем нулевые приближения $q_1^{(0)}, q_2^{(0)}, \dots, q_{p'}^{(0)}$ для столбцов $q_1, q_2, \dots, q_{p'}$ матрицы Q .

Затем определяем следующее (первое) приближение $V^{(1)}$ и т. д.

Специальный прием определения столбцов q_i ($i = 1, 2, \dots, p'$) матрицы Q при известной матрице $\Psi = QQ^T$ опирается на то, что матрица Ψ может быть представлена в виде $\Psi = q_1 q_1^T + q_2 q_2^T + \dots + q_{p'} q_{p'}^T$. Используя специфику выбранных идентифицирующих условий, определяют вначале столбец q_1 . Затем переходят к матрице $\Psi_1 = \Psi - q_1 q_1^T = q_2 q_2^T + \dots + q_{p'} q_{p'}^T$ и определяют столбец q_2 и т. д.

Статистическое оценивание факторных нагрузок q_{ij} и остаточных дисперсий v_{ii} . Оценивание производится либо методом максимального правдоподобия (см. гл. 7), либо так называемым *центроидным методом*. Первый метод используется обычно при идентифицирующих условиях типа 1) и 2). Он хотя и дает эффективные оценки для q_{ij} и v_{ii} , но требует постулирования закона распределения исследуемых величин (разработан лишь в нормальном случае), а также весьма обременительных вычислений. Центроидный метод используется при идентифицирующих условиях типа 3). Давая оценки, близкие к оценкам максимального правдоподобия, он, как и всякий непараметрический метод, является более «устойчивым» по отношению к отклонениям от нормальности исследуемых признаков и требует меньшего объема вычислений. Однако из-за определенного произвола в его процедуре, которая приведена ниже, статистическая оценка метода, исследование его выборочных свойств (в общем случае) практически невозможны.

Общая схема реализации метода максимального правдоподобия следующая. Составляется логарифмическая функция правдоподобия как функция неизвестных параметров q_{ij} и v_{ii} , отвечающая исследуемой модели, т. е. с учетом нормальности $X_1, \dots, X_n$ модели (13.26) и соответственно (13.27); в качестве дополнительных идентифицирующих условий берутся условия 1) или 2). С помощью дифференцирования этой функции правдоподобия по каждому из неизвестных параметров и приравнивания по-

лученных частных производных к нулю получается система уравнений, в которой известными величинами являются выборочные ковариации $\hat{\sigma}_{ij}$, а также числа p и p' , а неизвестными — искомые параметры q_{ij} и v_{ii} . И наконец, предлагается вычислительная (как правило, итерационная) процедура решения этой системы. Реализация описанной выше (для случаев 1 и 2) общей итерационной вычислительной схемы с заменой неизвестной ковариационной матрицы исходных признаков Σ ее выборочным аналогом $\hat{\Sigma}$ приведет как раз к оценкам максимального правдоподобия параметров q_{ij} и v_{ii} ($i = 1, 2, \dots, p; j = 1, 2, \dots, p'$). Отметим также, что при достаточно общих ограничениях доказана асимптотическая нормальность оценок максимального правдоподобия $\hat{Q}$ и $\hat{V}$, что дает основу для построения соответствующих интервальных оценок.

Как выше отмечено, центроидный метод является одним из способов реализации вычислительной схемы, приспособленной для выявления структуры модели факторного анализа и оценки неизвестных параметров в случае идентифицирующих условий типа 3). Этот метод поддается весьма простой геометрической интерпретации. отождествим исследуемые признаки $x^{(1)}, \dots, x^{(p)}$ с векторами, выходящими из начала координат некоторого вспомогательного p -мерного пространства, построенными таким образом, чтобы косинусы углов между $x^{(i)}$ и $x^{(j)}$ равнялись бы их парным корреляциям (r_{ij}), а длины векторов $x^{(i)}$ — стандартным отклонениям соответствующих переменных ($\sigma_{ii}^{1/2}$). Далее изменим, если необходимо, направления, т. е. знаки отдельных векторов так, чтобы как можно больше корреляций стало положительными. Тогда векторы будут иметь тенденцию к группировке в одном направлении в пучок. После этого первый общий фактор $f^{(1)}$ определяется как нормированная (т. е. как вектор единичной длины) сумма всех исходных векторов пучка, и, следовательно, он будет проходить каким-то образом через середину (центр) этого пучка; отсюда название «центроид» для общего фактора в этом случае.

Переходя затем к остаточным переменным $x^{(i1)} = x^{(i)} - q_{i1}f^{(1)}$, подсчитывая ковариационную матрицу $\Sigma^{(1)} = \Sigma - q_1q_1^T$ для этих остаточных переменных и проделывая относительно $x^{(i1)}$ и $\Sigma^{(1)}$ ту же самую процедуру построения пучка и т. п., выделяем второй общий фактор («второй центроид») $f^{(2)}$ и т. д.

Формализация этих соображений приводит к следующей итерационной схеме вычислений по определению факторных нагрузок q_{ij} и остаточных дисперсий v_{ii} . Задаемся некоторым начальным приближением $V^{(0)}$

для дисперсий остатков V . Обычно полагают

$$v_{ii}^{(0)} = \hat{\sigma}_{ii} \left[1 - \max_{\substack{1 \leq j \leq p \\ (j \neq i)}} |r_{ji}| \right].$$

Подсчитываем $\Psi^{(0)} = \hat{\Sigma} - V^{(0)}$. Выбираем в качестве нулевого приближения $b_1^{(0)}$ первого столбца b_1 вспомогательной матрицы B столбец, состоящий из одних единиц

$$b_1^{(0)} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Далее определяем нулевое приближение $q_1^{(0)}$ первого столбца матрицы нагрузок

$$q_1^{(0)} = \frac{\Psi^{(0)} b_1^{(0)}}{(b_1^{(0)\top} \Psi^{(0)} b_1^{(0)})^{\frac{1}{2}}}.$$

Затем вычисляется матрица $\Psi_1^{(0)} = \Psi^{(0)} - q_1^{(0)} q_1^{(0)\top}$ и определяется нулевое приближение $q_2^{(0)}$ второго столбца матрицы нагрузок

$$q_2^{(0)} = \frac{\Psi_1^{(0)} b_2^{(0)}}{(b_2^{(0)\top} \Psi_1^{(0)} b_2^{(0)})^{\frac{1}{2}}}, \quad (13.30)$$

где вектор $b_2^{(0)}$ состоит только из $+1$ или -1 , а знаки подбираются из условия максимизации знаменателя правой части (13.30) и т. д. Получив, таким образом, нулевое приближение $Q^{(0)} = (q_1^{(0)}, \dots, q_{p'}^{(0)})$ для матрицы нагрузок Q , вычисляем $V^{(1)} = \Sigma - Q^{(0)} Q^{(0)\top}$ и переходим к следующей итерации. При этом матрица $B^{(1)}$ не обязана совпадать с $B^{(0)}$. Кстати, как нетрудно усмотреть из вышесказанного, i -й столбец матрицы B задает веса, с которыми суммируются векторы одного пучка для образования i -го общего фактора («центроида»). Поскольку смысл центроидной процедуры в простом суммировании векторов пучка, она иногда так и называется — «процедура простого суммирования», то исследователю остается определить лишь нужное направление каждого из векторов пучка, т. е. знаки единиц, образующих столбцы b_i . Непосредственная ориентация (при подборе знаков у компонент вектора b_i) на максимизацию выражений $b_i^{(\nu)\top} \Psi_{i-1}^{(\nu)} b_i^{(\nu)}$ хотя и несколько сложнее реализуема, чем некоторые эвристические приемы, опирающиеся на анализ знаков элементов остаточных матриц Ψ_{i-1} , но быстрее и надежнее приводит к выделению именно таких центроидов, которые при заданном p' будут обуславливать

возможно большую часть общей дисперсии исходных признаков, т.е. минимизировать дисперсию остаточных компонент u_i .

Недостатком центроидного метода является зависимость центроидных нагрузок q_{ij} от шкалы, в которой измерены исходные признаки. Поэтому исходные признаки $x^{(i)}$ обычно нормируют с помощью среднеквадратических отклонений $\sigma_{ii}^{1/2}$, так что выборочная ковариационная матрица $\hat{\Sigma}$ заменяется во всех рассуждениях выборочной корреляционной матрицей $\hat{R}$.

Анализируя описанную выше процедуру центроидного метода, нетрудно понять, что построенные таким способом общие факторы могут интерпретироваться как первые p' «условных» главных компонент матрицы $\hat{\Sigma} - V$, найденные при дополнительном условии, что компоненты соответствующих собственных векторов могут принимать лишь два значения: плюс или минус 1.

Оценка значений общих факторов. Это одна из основных задач исследования. Действительно, мало установить лишь сам факт существования небольшого числа скрыто действующих общих факторов $f^{(1)}, \dots, f^{(p')}$, объясняющих природу взаимной коррелированности исходных признаков и основную часть их дисперсии. Желательно непосредственно определить эти общие факторы, описать их в терминах исходных признаков и постараться дать им удобную содержательную интерпретацию.

Приведем здесь идеи и результаты двух распространенных методов решения этой задачи, предложенных в разное время М. Бартлеттом (1938 г.) и Г. Томсоном (1951 г.). В обоих случаях предполагаем задачу статистического оценивания неизвестных нагрузок $\hat{Q}' = (\hat{q}_{ij})$ и остаточных дисперсий $\hat{V} = (\hat{v}_{ii})$ уже решенной.

Метод Бартлетта рассматривает отдельно для каждого фиксированного номера наблюдения ν ($\nu = 1, 2, \dots, n$) модель (13.25) как регрессию признака x_ν по аргументам $\hat{q}_{\cdot 1}, \hat{q}_{\cdot 2}, \dots, \hat{q}_{\cdot p'}$; при этом верхний индекс $i = 1, 2, \dots, p$ у признака (и соответствующий первый нижний индекс у нагрузок) играет в данном случае роль номера наблюдений в этой регрессионной схеме, так что

$$x_\nu^{(i)} = \sum_{j=1}^{p'} f_\nu^{(j)} q_{ij} + u_\nu^{(i)} \quad (i = 1, \dots, p).$$

Таким образом, величины $f_\nu^{(1)}, f_\nu^{(2)}, \dots, f_\nu^{(p')}$ интерпретируются как неизвестные коэффициенты регрессии x_ν по $\hat{q}_{\cdot 1}, \hat{q}_{\cdot 2}, \dots, \hat{q}_{\cdot p'}$. В соответ-

ствии с известной техникой метода наименьших квадратов (с учетом «неравноточности» измерений, т. е. того, что, вообще говоря, $\mathbf{D} x^{(i_1)} \neq \mathbf{D} x^{(i_2)}$ при $i_1 \neq i_2$), определяющей неизвестные коэффициенты регрессии $\hat{F}_\nu = (\hat{f}_\nu^{(1)}, \dots, \hat{f}_\nu^{(p')})^\top$ из условия

$$\sum_{i=1}^p \frac{1}{\sigma_{ii}} \left(x_\nu^{(i)} - \sum_{j=1}^{p'} \hat{f}_\nu^{(j)} \hat{q}_{ij} \right)^2 = \min_{F_\nu} \sum_{i=1}^p \frac{1}{\sigma_{ii}} \times \left(x_\nu^{(i)} - \sum_{j=1}^{p'} f_\nu^{(j)} \hat{q}_{ij} \right)^2,$$

получаем

$$\hat{F}_\nu = (\hat{\mathbf{Q}}^\top \hat{\mathbf{V}}^{-1} \hat{\mathbf{Q}})^{-1} \hat{\mathbf{Q}}^\top \hat{\mathbf{V}}^{-1} X_\nu \quad (\nu = 1, \dots, n). \quad (13.31)$$

Очевидно, если исследуемый вектор наблюдений X нормален, то эти оценки являются одновременно и оценками максимального правдоподобия. Нестрогость данного метода — в замене истинных (неизвестных нам) величин q_{ij} и v_{ii} их приближенными (оценочными) значениями $\hat{q}_{ij}$ и $\hat{v}_{ii}$.

Модель Томсона рассматривает модель (13.26) как бы «вывернутой наизнанку», а именно как регрессию зависимых переменных $f^{(1)}, \dots, f^{(p')}$ по аргументам $x^{(1)}, \dots, x^{(p)}$. Тогда коэффициенты $\hat{c}_{ij}$ в соотношениях

$$\hat{f}^{(i)} = \sum_{j=1}^p \hat{c}_{ij} x^{(j)} \quad (i = 1, \dots, p'),$$

или в матричной записи

$$\hat{\mathbf{F}} = \mathbf{C} X,$$

где $\mathbf{C}$ — матрица коэффициентов c_{ij} размера $p' \times p$, находят в соответствии с методом наименьших квадратов из условия

$$\sum_{\nu=1}^n \sum_{i=1}^{p'} \left(\hat{f}_\nu^{(i)} - \sum_{j=1}^p \hat{c}_{ij} x_\nu^{(j)} \right)^2 = \min_{c_{ij}} \sum_{\nu=1}^n \sum_{i=1}^{p'} \left(\hat{f}_\nu^{(i)} - \sum_{j=1}^p c_{ij} x_\nu^{(j)} \right)^2. \quad (13.32)$$

Поскольку решение экстремальной задачи (13.32) выписывается в терминах ковариаций $x^{(i)}$ и $f^{(j)}$, то отсутствие наблюдений по зависимым переменным $f^{(j)}$ можно компенсировать знанием этих ковариаций,

так как легко подсчитать, что

$$E \left\{ \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(p)} \\ f^{(1)} \\ \vdots \\ f^{(p')} \end{bmatrix} (x^{(1)}, \dots, x^{(p)}, f^{(1)}, \dots, f^{(p')}) \right\} = \begin{pmatrix} \mathbf{Q}\mathbf{Q}^\top + \mathbf{V} & \mathbf{Q} \\ \mathbf{Q}^\top & \mathbf{I}_{p'} \end{pmatrix}.$$

Отсюда, используя известные формулы метода наименьших квадратов (см. гл. 2 тома 2), получаем (с заменой матриц $\mathbf{Q}$ и $\mathbf{V}$ их выборочными аналогами)

$$F_\nu = (\mathbf{I} + \mathbf{\Gamma})^{-1} \hat{\mathbf{Q}}^\top \hat{\mathbf{V}}^{-1} X_\nu \quad (\nu = 1, 2, \dots, n), \quad (13.33)$$

где матрица $\mathbf{\Gamma}$ (размера $p \times p$) определяется соотношением

$$\mathbf{\Gamma} = \hat{\mathbf{Q}}^\top \hat{\mathbf{V}}^{-1} \hat{\mathbf{Q}}.$$

Сравнение выражений (13.31) и (13.33) позволяет получить явное соотношение между решениями по методу Бартлетта $\hat{F}^{(B)}$ и методом Томсона $\hat{F}^{(T)}$

$$\hat{F}^{(B)} = (\mathbf{I} + \mathbf{\Gamma}^{-1}) \hat{F}^{(T)}.$$

Если элементы матрицы $\hat{\mathbf{Q}}' \hat{\mathbf{V}}^{-1} \hat{\mathbf{Q}}$ достаточно велики, то эти два метода будут давать близкие решения.

Статистическая проверка гипотез. Проверка гипотез, связанных с природой и параметрами используемой модели факторного анализа, составляет один из необходимых моментов исследования. Теория статистических критериев применительно к моделям факторного анализа разработана весьма слабо. Пока удалось построить лишь так называемые критерии адекватности модели, т. е. критерии, предназначенные для проверки гипотезы H_0 , заключающейся в том, что исследуемый вектор наблюдений X допускает представление с помощью модели факторного анализа (13.26) с данным (заранее выбранным) числом общих факторов p' . При этом критическая статистика $\gamma(X_1, \dots, X_n)$, т. е. функция от результатов наблюдения, по значению которой принимается решение об отклонении или непротиворечивости высказанной гипотезы H_0 , зависит от вида дополнительных (идентифицирующих) условий модели. Так если рассматривается модель с дополнительными идентифицирующими условиями вида 1), т. е. дополнительно постулируются диагональность матрицы $\mathbf{\Gamma} = \hat{\mathbf{Q}}^\top \hat{\mathbf{V}}^{-1} \hat{\mathbf{Q}}$, то гипотеза H_0 отвергается (с вероятностью ошибиться,

приблизительно равной α) в случае

$$\gamma_1(X_1, \dots, X_n) = n(\ln |\hat{\mathbf{V}}| + \ln |\mathbf{I} + \mathbf{\Gamma}| - \ln |\hat{\mathbf{\Sigma}}|) > \chi^2_\alpha(\nu_1),$$

где число степеней свободы $\nu_1 = \frac{1}{2}[(p-p')^2 - (p+p')]$; его положительность обеспечивается условием $p' < (p-1)/2$, а $\chi^2_\alpha(\nu_1)$ — как и ранее, величина 100 α %-ной точки χ^2 -распределения с ν_1 степенями свободы (находится из таблиц).

На языке ковариационных матриц гипотеза H_0 означает в данном случае, что элементы матрицы $\hat{\mathbf{\Sigma}} - (\hat{\mathbf{Q}}\hat{\mathbf{Q}}^T + \hat{\mathbf{V}})$ должны лишь статистически незначимо отличаться от нуля, или, что эквивалентно, матрица $\hat{\mathbf{\Sigma}} - \hat{\mathbf{V}}$ должна иметь ранг, равный p' . А это в свою очередь означает, что последние $p - p'$ характеристических корней $\hat{\lambda}_{p'+1}, \dots, \hat{\lambda}_p$ уравнения $|\hat{\mathbf{\Sigma}} - \hat{\mathbf{V}} - \lambda \hat{\mathbf{V}}| = 0$ должны лишь незначимо отличаться от нуля. Статистика $\gamma_1(X_1, \dots, X_n)$ может быть записана в терминах этих характеристических корней:

$$\gamma_1(X_1, \dots, X_n) = n \sum_{i=p'+1}^p \ln(1 + \hat{\lambda}_i).$$

Если же в качестве идентифицирующих условий дополнительно к (13.26), или, что то же, к (13.27), постулируется наличие какого-то заранее заданного числа m нулевых нагрузок q_{ij} из общего числа $p \cdot p'$ на определенных («специфических») позициях, то гипотеза H_0 отвергается (с вероятностью ошибиться, приблизительно равной α) в случае, когда

$$\gamma^2(X_1, \dots, X_n) = n(\ln |\hat{\mathbf{V}}|) + \ln |\hat{\mathbf{Q}}^T \hat{\mathbf{V}}^{-1} \hat{\mathbf{\Sigma}} \hat{\mathbf{V}}^{-1} \hat{\mathbf{Q}}| - \ln |\mathbf{\Gamma}| - \ln |\hat{\mathbf{\Sigma}}| > \chi^2_\alpha(\nu_2),$$

где число степеней свободы $\nu_2 = \frac{1}{2}p(p-1) - (pp' - m)$.

Иногда удобнее вычислять критическую статистику $\gamma_2(X_1, \dots, X_n)$ в терминах характеристических корней $\hat{z}_1, \hat{z}_2, \dots, \hat{z}_p$ (нумерованных в порядке убывания их величин) выборочной корреляционной матрицы $\mathbf{R}$ исследуемого вектора наблюдений X :

$$\begin{aligned} \gamma_2(X_1, \dots, X_n) = & \left(n - \frac{2p+11}{6} - \frac{2p'}{3} \right) \\ & \times \left[(p-p') \ln \left(\frac{\sum_{i=p'+1}^p \hat{z}_i}{p-p'} \right) - \sum_{i=p'+1}^p \ln \hat{z}_i \right]. \end{aligned}$$

Статистики $\gamma_1(X_1, \dots, X_n)$ и $\gamma_2(X_1, \dots, X_n)$ получены в результате реализации известной схемы критерия отношения правдоподобия.

До сих пор не удалось построить многомерной решающей процедуры типа $\hat{p}'(\hat{\Sigma})$, т.е. оценки для неизвестного числа общих факторов p' . В настоящее время приходится ограничиваться последовательной эксплуатацией критериев адекватности $H_0: p' = p'_0$ (p'_0 заранее задано) при альтернативе $H_1: p' > p'_0$. Если гипотеза H_0 отвергается, то переходят к проверке гипотезы $H'_0: p' = p'_0 + 1$ при альтернативе $H'_1: p' > p'_0 + 1$ и т.д. Однако по уровням значимости α каждой отдельной стадии такой процедуры трудно сколько-нибудь точно судить о свойствах всей последовательной процедуры в целом.

Пользуясь асимптотической нормальностью оценок $\hat{Q}$ и $\hat{V}$, можно было бы попытаться строить критерии для проверки гипотез, касающихся значений факторных нагрузок, например, гипотез о том, что некоторые признаки не зависят от заранее определенных факторов, т.е. что на определенных местах матрицы $\hat{Q}$ стоят элементы, статистически незначимо отличающиеся от нуля. Однако построение этих критериев затруднено из-за сложности процедуры вычисления ковариационных матриц оценок $\hat{Q}$ и $\hat{V}$.

Продемонстрируем реализацию некоторых этапов факторного анализа на примере.

Пример 13.4. В табл. 13.3 приведены коэффициенты корреляции между отметками по шести школьным предметам, подсчитанные по выборке из 220 учащихся.

Таблица 13.3

Содержательный смысл признака	Номер признака						Нагрузка факторов на признаки	
	1	2	3	4	5	6	q_{i1}	q_{i2}
Отметка по галльскому языку $x^{(1)}$	1	0,439	0,410	0,288	0,329	0,248	0,606	0,337
английскому языку $x^{(2)}$	0,439	1	0,351	0,354	0,320	0,329	0,611	0,197
истории $x^{(3)}$	0,410	0,351	1	0,164	0,190	0,181	0,458	0,384
арифметике $x^{(4)}$	0,288	0,354	0,161	1	0,595	0,570	0,683	-0,365
алгебре $x^{(5)}$	0,329	0,320	0,190	0,595	1	0,464	0,686	-0,335
геометрии $x^{(6)}$	0,248	0,329	0,181	0,470	0,464	1	0,575	-0,212

В последних двух столбцах таблицы даны факторные нагрузки q_{i1} , q_{i2} на исследуемые признаки в бифакторной модели ($p = 2$), подсчитанные

по приведенной здесь корреляционной матрице с помощью центроидного метода. Простой анализ величин и знаков этих нагрузок склоняет нас к тому, чтобы интерпретировать первый фактор $f^{(1)}$ как фактор общей одаренности, а второй фактор $f^{(2)}$ — как фактор гуманитарной одаренности.

Метод Г. Томсона (13.33) дает в качестве оценки общих факторов выражения:

$$\begin{aligned}\hat{f}^{(1)} &= 0,245x^{(1)} + 0,208x^{(2)} + 0,158x^{(3)} + 0,278x^{(4)} + 0,271x^{(5)} + 0,157x^{(6)}; \\ \hat{f}^{(2)} &= 0,352x^{(1)} + 0,201x^{(2)} + 0,309x^{(3)} - 0,351x^{(4)} - 0,303x^{(5)} - 0,126x^{(6)}.\end{aligned}$$

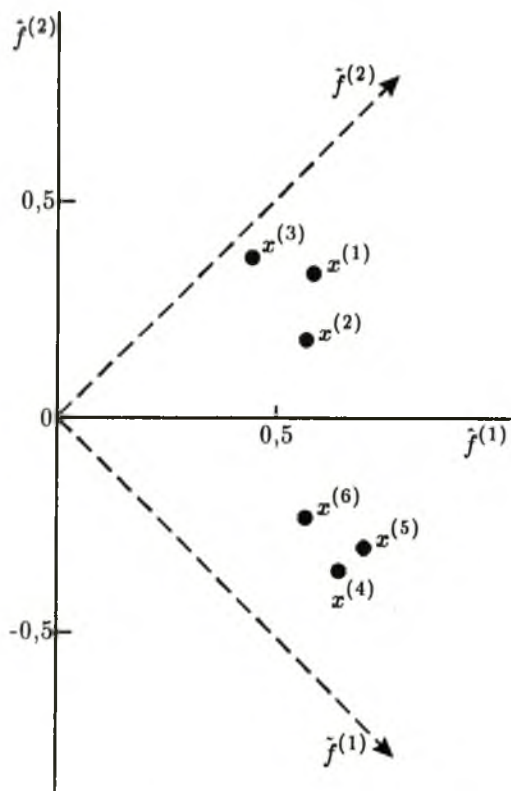


Рис. 13.3. Изображение исходных признаков $x^{(1)}, \dots, x^{(6)}$ в плоскости двух общих факторов $\hat{f}^{(1)}, \hat{f}^{(2)}$

В целях геометрической интерпретации центроидного метода рассмотрим рис. 13.3, на котором осями координат являются общие факторы $\hat{f}^{(1)}$ и $\hat{f}^{(2)}$, а координаты точек $(\hat{f}_i^{(1)}, \hat{f}_i^{(2)}) = (q_{i1}, q_{i2})$ определяются нагрузками

ми i -го исходного признака на общие факторы ($i = 1, 2, \dots, 6$). Соответственно точку (q_{11}, q_{12}) удобно интерпретировать как изображение i -го исходного признака $x^{(i)}$. Расположение точек на рис. 13.3 свидетельствует о естественном распадении совокупности исходных признаков на две группы: группу гуманитарных признаков $(x^{(1)}, x^{(2)}, x^{(3)})$ и группу математических признаков $(x^{(4)}, x^{(5)}, x^{(6)})$.

Подобная геометрическая интерпретация помогает выбрать вращение системы общих факторов, *наиболее подходящее в отношении возможности их содержательной интерпретации*. Дело в том, что, как уже отмечали, параметры модели факторного анализа, в том числе и сами общие факторы $f^{(1)}, f^{(2)}, \dots, f^{(p')}$, определяются не однозначно, а лишь с точностью до некоторого ортогонального преобразования, т.е. с точностью до вращения осей $f^{(1)}, f^{(2)}, \dots, f^{(p')}$ в пространстве. При этом выбор окончательного решения, т.е. закрепление системы $f^{(1)}, f^{(2)}, \dots, f^{(p')}$ в определенном положении, находится в распоряжении исследователя. Другими словами, исследователь должен решить вопрос: как, располагая некоторым частным решением $f^{(1)}, f^{(2)}, \dots, f^{(p')}$, полученным, например, с помощью центроидного метода, выбрать такое ортогональное преобразование, такой поворот осей $f^{(1)}, f^{(2)}, \dots, f^{(p')}$, при котором получаемые при этом новые общие факторы $\tilde{f}^{(1)}, \tilde{f}^{(2)}, \dots, \tilde{f}^{(p')}$ допускают наиболее естественную и убедительную интерпретацию. Рассматривая расположение исходных признаков в плоскости $\tilde{f}^{(1)} \tilde{f}^{(2)}$ или в пространстве, натянутом на первые три общих фактора, естественно повернуть координатную систему таким образом, чтобы координатные оси прошли через наиболее четко выраженные сгущения точек-признаков (см. поворот, намеченный пунктирными осями $\tilde{f}^{(1)}$ и $\tilde{f}^{(2)}$ на рис. 13.3). При этом иногда бывает полезно отказаться от ортогональности общих факторов, переходя к косоугольной системе координат.

13.4. Некоторые эвристические методы снижения размерности

13.4.1. Природа эвристических методов

Описанные выше сокращения размерности исследуемого признакового пространства (метод главных компонент и модели факторного анализа) допускали интерпретацию в терминах той или иной строгой вероятност-

ной модели и, следовательно, подразумевали возможность исследования свойств рассматриваемых процедур в рамках теории математической статистики. В данном пункте речь пойдет о методах, подчиненных некоторым частным целевым установкам (наименьшее искажение геометрической структуры исходных «выборочных точек», наименьшее искажение их эталонного разбиения на классы и т. д.), но не формулируемых в терминах вероятностно-статистической теории¹. Процедура выбора целевой установки, подходящей именно для данной конкретной задачи, практически не формализована, носит эвристический характер, т. е., как правило, обуславливается лишь опытом и интуицией исследователя. Поэтому будем называть такие методы *эвристическими*.

При отсутствии априорной или выборочной предварительной информации о природе исследуемого вектора наблюдений и о генеральных совокупностях, из которых эти наблюдения извлекаются, точно в таком же невыгодном положении находятся и методы факторного анализа и главных компонент. Однако для них все-таки существует принципиальная возможность теоретического обоснования (при наличии соответствующей дополнительной информации), в то время как лишь некоторые из эвристических методов удастся впоследствии теоретически обосновать в рамках строгой математической модели.

Подчеркнем, что факт описания здесь методов снижения размерности, не использующих предварительной информации, например, обучающих и квазиобучающих выборок, целесообразно расценивать лишь как следствие признания неизбежности ситуаций, в которых такой информации не имеется, но не как стремление рекламировать эти методы в качестве наиболее эффективных. В действительности же обоснование и эффективное решение задач снижения размерности без слепой надежды на удачу можно, по нашему мнению, получить лишь на пути глубокого профессионального анализа, дополненного статистическими методами, использующими предварительную выборочную (обучающую) информацию.

13.4.2. Метод экстремальной группировки признаков

При изучении сложных объектов, заданных многими параметрами, возникает задача разбиения параметров на группы, каждая из которых

¹ Отсутствие строгой вероятностно-статистической модели, лежащей в основе тех или иных методов, не исключает возможности использования отдельных вероятностно-статистических понятий и соответствующей терминологии, как это имеет место, например, в методе экстремальной группировки признаков, в методе корреляционных плеяд и некоторых других.

характеризует объект с какой-либо одной стороны. Но получение легко интерпретируемых результатов осложняется тем, что во многих приложениях измеряемые параметры (признаки) лишь косвенно отражают существенные свойства, которыми характеризуется данный объект.

Так, в психологии измеряемые параметры — это реакции людей на различные тесты, а выражением существенных свойств, общими факторами являются такие характеристики, как тип нервной системы, работоспособность и т. п. Подобная природа формирования набора частных характеристик объекта или системы присуща широкому классу явлений и процессов в экономике, социологии, медицине, педагогике и т. д.

Оказывается, что во многих случаях изменение какого-либо общего фактора сказывается неодинаково на измеряемых признаках, в частности, исходная совокупность из p признаков обнаруживает такое естественное «расщепление» на сравнительно (с p) небольшое количество групп, при котором изменение признаков, относящихся к какой-либо одной группе, обуславливается в основном каким-то одним общим фактором, своим для каждой такой группы. После принятия этой гипотезы разбиение на группы естественно строить так, чтобы параметры, принадлежащие к одной группе, были коррелированы сравнительно сильно, а параметры, принадлежащие к разным группам, — слабо. После такого разбиения для каждой группы признаков строится случайная величина, которая в некотором смысле наиболее сильно коррелирована с параметрами данной группы; эта случайная величина интерпретируется как искомый фактор, от которого существенно зависят все параметры данной группы.

Очевидно, подобная схема является одним из частных случаев общей логической схемы факторного анализа. В отличие от ранее описанных классических моделей факторного анализа при эвристически-оптимизационном подходе группировка признаков и выделение общих факторов делаются на основе экстремизации некоторых эвристически введенных функционалов. Разбиения, оптимизирующие функционал J_1 или J_2 (см. ниже), называются *экстремальной группировкой параметров*. Вообще под задачей экстремальной группировки набора случайных величин $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ на заранее заданное число классов p' понимают отыскание такого набора подмножеств $S_1, S_2, \dots, S_{p'}$ натурального ряда чисел $1, 2, \dots, p$, что $\bigcup_{l=1}^{p'} S_l = \{1, 2, \dots, p\}$, а $S_l \cap S_q = \emptyset$ при $l \neq q$, и таких p' нормированных (т. е. с единичной дисперсией $D f^{(i)} = 1$) факторов $f^{(1)}, f^{(2)}, \dots, f^{(p')}$, которые максимизируют какой-либо критерий оптимальности.

Остановимся здесь на алгоритмах для двух различных критериев оптимальности.

Первый алгоритм экстремальной группировки признаков в качестве критерия оптимальности использует функционал

$$J_1 = \sum_{i \in S_1} [\text{согг}(x^{(i)}, f^{(1)})]^2 + \dots + \sum_{i \in S_{p'}} [\text{согг}(x^{(i)}, f^{(p')})]^2,$$

в котором под $\text{согг}(x, f)$ понимается обычный парный коэффициент корреляции между признаком x и фактором f . Обозначим $A_l = \{x^{(i)}, i \in S_l\}$, $l = 1, 2, \dots, p'$. Максимизация функционала J_1 (как по разбиению признаков на группы $A_1, \dots, A_{p'}$, так и по выбору факторов $f^{(1)}, f^{(2)}, \dots, f^{(p')}$) отвечает требованию такого разбиения параметров, когда в одной группе оказываются наиболее «близкие» между собой, в смысле степени коррелированности, признаки: в самом деле, при максимизации функционала J_1 , для каждого фиксированного набора случайных величин $f^{(1)}, f^{(2)}, \dots, f^{(p')}$, в одну l -ю группу будут попадать такие признаки, которые наиболее сильно коррелированы с величиной $f^{(l)}$; в то же время среди всех возможных наборов случайных величин $f^{(1)}, f^{(2)}, \dots, f^{(p')}$ будет выбираться такой набор, что каждая из величин $f^{(l)}$ в среднем наиболее «близка» ко всем признакам своей группы.

Очевидно, что при заданных классах $S_1, S_2, \dots, S_{p'}$ оптимальный набор факторов $f^{(1)}, f^{(2)}, \dots, f^{(p')}$ получается в результате независимой максимизации каждого слагаемого

$$\sum_{i \in S_l} [\text{согг}(x^{(i)}, f^{(l)})]^2 \quad (l = \overline{1, p'}),$$

откуда

$$\max_{f^{(1)}, f^{(2)}, \dots, f^{(l)}} J_1 = \sum_{l=1}^{p'} \lambda_l^2,$$

где λ_l — максимальное собственное значение матрицы $\mathbf{R}_l$, составленной из коэффициентов корреляции переменных, входящих в A_l . При этом оптимальный набор факторов $f^{(l)}$, $l = 1, 2, \dots, p'$, задается формулами:

$$f^{(l)} = \frac{\sum_{i \in S_l} \alpha_i^{(l)} x^{(i)}}{\sqrt{\sum_{i, j \in S_l} \alpha_i^{(l)} \alpha_j^{(l)} r_{ij}}}, \quad l = 1, 2, \dots, p', \quad (13.34)$$

где $r_{ij} = \text{согг}(x^{(i)}, x^{(j)})$, а $\alpha^{(l)} = (\alpha_1^{(l)}, \alpha_2^{(l)}, \dots, \alpha_{m_l}^{(l)})$ — собственный вектор матрицы $\mathbf{R}_l$, отвечающий максимальному собственному значению λ_l , т. е. $\mathbf{R}_l \cdot \alpha^{(l)} = \lambda_l \cdot \alpha^{(l)}$.

С другой стороны, считая известными факторы $f^{(1)}, f^{(2)}, \dots, f^{(p')}$, не трудно построить разбиение $S_1, S_2, \dots, S_{p'}$, максимизирующее J_1 при фиксированных $f^{(1)}, f^{(2)}, \dots, f^{(p')}$, а именно:

$$S_l = \left\{ i: \text{согг}^2(x^{(i)}, f^{(l)}) \geq \text{согг}^2(x^{(i)}, f^{(q)}) \text{ для всех } q = 1, 2, \dots, p' \right\}. \quad (13.35)$$

Соотношения (13.34) и (13.35) являются необходимыми условиями максимума J_1 .

Для одновременного нахождения оптимального разбиения $S_1, S_2, \dots, S_{p'}$ и оптимального набора факторов $f^{(1)}, f^{(2)}, \dots, f^{(p')}$ предлагается итерационный алгоритм, последовательно осуществляющий выбор оптимальных (по отношению к разбиению, полученному на предыдущем шаге) факторов, а затем выбор разбиения, оптимального к факторам, полученным на предыдущем шаге.

Пусть на ν -м шаге итерации построено разбиение параметров на группы $A_1, \dots, A_{p'}$. Для каждой такой группы параметров строят факторы $f_\nu^{(i)}$ по формуле (13.34) и новое $(\nu + 1)$ разбиение параметров $A_1^{(\nu+1)}, \dots, A_{p'}^{(\nu+1)}$ в соответствии с правилом: параметр $x^{(i)}$ относится к группе $A_i^{(\nu+1)}$, если

$$\text{согг}^2(x^{(i)}, f_\nu^{(i)}) \geq \text{согг}^2(x^{(i)}, f_\nu^{(q)}) \quad (q = 1, 2, \dots, p'). \quad (13.36)$$

Если для некоторого параметра $x^{(i)}$ найдутся два или более факторов таких, что для $x^{(i)}$ и этих факторов в (13.36) имеет место равенство, то параметр $x^{(i)}$ относится к одной из соответствующих групп произвольно.

Очевидно, что на каждом шаге итераций функционал J_1 не убывает, поэтому данный алгоритм будет сходиться к максимуму. Максимум может быть локальным.

Для описания второго алгоритма экстремальной группировки признаков введем функционал

$$J_2 = \sum_{i \in S_1} |\text{согг}(x^{(i)}, f^{(1)})| + \sum_{i \in S_2} |\text{согг}(x^{(i)}, f^{(2)})| \\ + \dots + \sum_{i \in S_{p'}} |\text{согг}(x^{(i)}, f^{(p')})|.$$

В содержательном смысле функционал J_2 похож на функционал J_1 и его максимизация также соответствует основному требованию к характеру разбиения признаков на группы. Доказано, что необходимыми и достаточными условиями максимума функционала J_2 являются следующие:

• разбиение параметров на группы $A_1, \dots, A_{p'}$ таково, что функционал

$$J_3 = \sum_{l=1}^{p'} \sqrt{D \left(\sum_{i \in S_l} g_i x^{(i)} \right)}$$

(где g_i — некоторые числовые коэффициенты, равные либо $+1$, либо -1) достигает максимума как по разбиению на группы, так и по значениям коэффициентов g_i . Здесь под Dz понимается, как обычно, дисперсия случайной величины z ;

• факторы $f^{(l)}$ определяются соотношениями

$$f^{(l)} = \frac{\sum_{i \in S_l} g_i x^{(i)}}{\sqrt{\sum_{i,j \in S_l} g_i g_j r_{ij}}} \quad (13.37)$$

Логическая схема доказательства этого следующая. Сначала, варьируя функционал J_2 и используя метод множителей Лагранжа для учета условия $D f^{(l)} = 1$, показывают, что в точке максимума функционала J_2 фактор $f^{(l)}$ имеет вид (13.37). Затем доказывается, что если $f^{(l)}$ имеет вид (13.37), то при любом наборе коэффициентов $g_i = \pm 1$ и любом разбиении параметров на группы имеет место соотношение $J_2 \geq J_3$, а если же J_3 достигает максимума, то $J_2 = J_3$. Из этого утверждения следует, в частности, что для нахождения групп S_l и факторов $f^{(l)}$ достаточно максимизировать функционал J_3 . При фиксированном разбиении на группы функционал J_3 достигает максимума тогда, когда для каждого l соответствующие коэффициенты g_i максимизируют величину

$$D \left(\sum_{i \in S_l} g_i x^{(i)} \right). \quad (13.38)$$

Поэтому естественно воспользоваться рекуррентной процедурой максимизации J_3 . В процедуре циклически перебираются переменные $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, на каждом шаге принимается решение об отнесении очередного параметра $x^{(j)}$ к одной из групп $A_1, \dots, A_{p'}$ и определяется знак g_i .

Пусть к ν -му шагу алгоритма построены разбиения параметров на группы $A_1^{(\nu)}, \dots, A_{p'}^{(\nu)}$, вычислены коэффициенты $g_1^{(\nu)}, g_2^{(\nu)}, \dots, g_{p'}^{(\nu)}$, равные $+1$ или -1 , и пусть на этом шаге рассматривается признак $x^{(i)} \in A_l^{(\nu)}$. Тогда строятся p' вспомогательных коэффициентов $g_{i,l}^{(\nu+1)}$ ($l = 1, \dots, p'$)

по формуле

$$g_{i,l}^{(\nu+1)} = \text{sign} \sum_{x^{(j)} \in A_l^{(\nu)}} g_j^{(\nu)} r_{ij} \quad (j \neq i),$$

где

$$\text{sign } x = \begin{cases} 1 & \text{при } x > 0, \\ 0 & \text{при } x = 0, \\ -1 & \text{при } x < 0 \end{cases}$$

и для всех $l = 1, 2, \dots, p'$ вычисляются разности

$$\Delta_l^{(\nu+1)} = \sqrt{D \left(\sum_{\substack{x_j \in S_l^{(\nu)} \\ j \neq i}} g_j^{(\nu)} x^{(j)} + g_{i,l}^{(\nu+1)} x^{(i)} \right)} - \sqrt{D \left(\sum_{\substack{x_j \in S_l^{(\nu)} \\ j \neq i}} g_j^{(\nu)} x^{(j)} \right)^2}.$$

Затем выбирается такой номер $l = l^*$, что

$$\Delta_{l^*}^{(\nu+1)} = \max_{1 \leq l \leq p'} \Delta_l^{(\nu+1)},$$

и признак $x^{(i)}$ исключается из группы A_l и присоединяется к группе A_{l^*} ; остальные группы признаков на этом шаге не меняются. В результате получаем новое разбиение признаков — $A_1^{(\nu+1)}, A_2^{(\nu+1)}, \dots, A_{p'}^{(\nu+1)}$. Новые значения коэффициентов $g_{i,l^*}^{(\nu+1)}$ определяются по формулам: $g_j^{(\nu+1)} = g_j^{(\nu)}$ (для $j \neq i$), $g_i^{(\nu+1)} = g_{i,l^*}^{(\nu+1)}$.

На следующем $(\nu + 1)$ -м шаге алгоритма рассматривается параметр $x^{(i+1)}$, если $i \neq p$, и $x^{(1)}$, если $i = p$.

Процедура заканчивается, если при рассмотрении всех признаков очередного цикла сохранились как разбиения признаков на группы, так и значения всех коэффициентов; полученное разбиение и значения коэффициентов рассматриваются как оптимальные (доказано, что на каждом шаге алгоритма значение J_3 не убывает).

Нетрудно проследить идейную близость метода экстремальной группировки факторов с методами, опирающимися на логическую схему факторного анализа. Так, например, отправляясь от общей модели вида (13.26)

$$x^{(i)} = \sum_{q=1}^{p'} l_{iq} f^{(q)} + u_i,$$

первую компоненту $f^{(1)}$ и «нагрузки» l_{i1} в методе главных компонент можно определять из условия минимума выражения $\sum_{i=1}^p D(x^{(i)} - l_{i1}f^{(1)})$ при нормирующем ограничении $D f^{(1)} = 1$. Решение этой условно экстремальной задачи очевидным образом сводится к нахождению максимума выражения $\sum_{i=1}^p [\text{corr}(x^{(i)}, f^{(1)})]^2$ при условии $D f^{(1)} = 1$.

Для построения следующего фактора $f^{(2)}$ (второй главной компоненты) рассматриваются случайные величины $x^{(i2)} = x^{(i)} - \text{corr}(x^{(i)}, f^{(1)})f^{(1)}$. Для этих случайных величин аналогичным образом находится свой фактор, который и является фактором $f^{(2)}$, и т. д.

Очевидно, что при реализации первого алгоритма метода экстремальной группировки признаков для каждой группы признаков A_l строится фактор, имеющий смысл первой главной компоненты для признаков этой группы.

В центроидном методе общий фактор ищут в виде

$$f^{(1)} = \sum_{i=1}^p g_i x^{(i)}, \quad (13.39)$$

где $g_i = \pm 1$ и g_i выбирается так, чтобы максимизировать величину

$$D f^{(1)} = D \left(\sum_{i=1}^p g_i x^{(i)} \right). \quad (13.40)$$

Сравнение выражений (13.39) и (13.40) с выражениями (13.37) и (13.38) показывает, что максимизация функционала J_2 приводит к построению для каждой группы признаков фактора, отличающегося на некоторый множитель от первого общего фактора, который был бы построен для этой группы центроидным методом.

13.4.3. Метод корреляционных плеяд

Задача разбиения признаков на группы часто имеет и самостоятельное значение. Например, в ботанике для систематизации вновь открытых растений делают разбиение набора признаков на группы так, чтобы 1-я группа характеризовала форму листа, 2-я группа — форму плода и т. д. В связи с этим и возник эвристический метод корреляционных плеяд.

Метод корреляционных плеяд, так же как и метод экстремальной группировки, предназначен для нахождения таких групп признаков — «плеяд», когда корреляционная связь, т. е. сумма модулей коэффициентов корреляции между параметрами одной группы (внутриплеядная связь)

достаточно велика, а связь между параметрами из разных групп (межплеядная) — мала. По определенному правилу по корреляционной матрице признаков образуют чертеж — граф, который затем с помощью различных приемов разбивают на подграфы. Элементы, соответствующие каждому из подграфов, и образуют плеяду.

Рассмотрим корреляционную матрицу $R = (r_{ij})$, $i, j = 1, 2, \dots, p$, исходных признаков. Нарисуем p кружков; внутри каждого кружка напишем номер одного из признаков. Каждый кружок соединяется линиями со всеми остальными кружками; над линией, соединяющей i -й и j -й элементы (ребром графа), ставится значение модуля коэффициента корреляции $|r_{ij}|$. Полученный таким образом чертеж рассматриваем как исходный граф.

Задавшись (произвольным образом или на основании предварительного изучения корреляционной матрицы) некоторыми пороговыми значениями коэффициента корреляции r_0 , исключаем из графа все ребра, которые соответствуют коэффициентам корреляции, по модулю меньшим r_0 . Затем задаем некоторое $r_1 > r_0$ и относительно него повторяем описанную процедуру. При некотором достаточно большом r граф распадается на несколько подграфов, т. е. таких групп кружков, что связи (ребра графа) между кружками различных групп отсутствуют. Очевидно, что для полученных таким образом плеяд внутриплеядные коэффициенты корреляции будут больше r , а межплеядные — меньше r .

В другом варианте корреляционных плеяд предлагается упорядочивать признаки и рассматривать только те коэффициенты корреляции, которые соответствуют связям между элементами в упорядоченной системе.

Упорядочение производится на основании принципа максимального корреляционного пути: все p признаков связываются при помощи $(p - 1)$ линий (ребер) так, чтобы сумма модулей коэффициентов корреляции была максимальной. Это достигается следующим образом: в корреляционной матрице находят наибольший по абсолютной величине коэффициент корреляции, например, $|r_{l,m}| = r^{(1)}$ (коэффициенты на главной диагонали матрицы, равные единице, не рассматриваются).

Рисуем кружки, соответствующие параметрам $x^{(l)}$ и $x^{(m)}$, и над «связью» между ними пишем значение $r^{(1)}$. Затем, исключив $r^{(1)}$, находим наибольший коэффициент в m -м столбце матрицы (это соответствует нахождению признака, который наиболее сильно после $x^{(l)}$ «связан» с $x^{(m)}$) и наибольший коэффициент в l -й строке матрицы (это соответствует нахождению признака, наиболее сильно после $x^{(m)}$ «связанного» с $x^{(l)}$). Из найденных таким образом двух коэффициентов корреляции выбирается наибольший — пусть это будет $|r_{ij}| = r^{(2)}$. Рисуем кружок $x^{(j)}$, соединяем

его с кружком $x^{(l)}$ и проставляем значение $r^{(2)}$. Затем находим признаки, наиболее связанные с $x^{(l)}$, $x^{(m)}$ и $x^{(j)}$, и выбираем из найденных коэффициентов корреляции наибольший. Пусть это будет $|r_{jq}| = r^{(3)}$. Требуем, чтобы на каждом шаге получался новый признак, поэтому признаки, уже изображенные на чертеже, исключаются, следовательно, $q \neq l$, $q \neq m$, $q \neq j$.

Далее рисуем кружок, соответствующий $x^{(q)}$, и соединяем его с $x^{(j)}$ и т. д. На каждом шаге находятся параметры, наиболее сильно связанные с двумя последними рассмотренными параметрами, а затем выбирается один из них, соответствующий большему коэффициенту корреляции. Процедура заканчивается после $(p - 1)$ -го шага; граф оказывается состоящим из p кружков, соединенных $(p - 1)$ ребром. Затем задается пороговое значение r , а все ребра, соответствующие меньшим, чем r , коэффициентам корреляции, исключаются из графа.

Назовем *незамкнутым графом* такой граф, для которого для любых двух кружков существует единственная траектория, составленная из линий связи, соединяющая эти два кружка. Очевидно, что во втором варианте метода корреляционных плеяд допускается построение только незамкнутых графов, а в первом варианте такое ограничение отсутствует. Поэтому разбиения на плеяды, полученные разными способами, могут не совпадать.

В работе [Лумельский В. Я.] приводятся результаты экспериментальной проверки алгоритмов экстремальной группировки параметров, а также сравнение полученных результатов с результатами, даваемыми методом корреляционных плеяд.

Эксперимент проводился на физиологическом материале: исследовались влияния шумов и вибрации на работоспособность и самочувствие. Регистрировались 33 признака ($p = 33$), из них 7 параметров, характеризующих температуру тела; 4 — кровяное давление; 14 — аудиометрию (порог слышимости на заданной частоте); 2 — дыхание; 4 — силу и выносливость рук и 2 (обособленных параметра) — пульс и скорость реакции.

С точки зрения физиолога «идеальным» было бы разбиение, при котором все характеристики температур образовали бы отдельную группу; параметры, характеризующие давление, — свою отдельную группу и т. д., обособленные параметры образовали бы группы, состоящие из одного элемента. Наиболее близким к «идеальному» оказалось разбиение, полученное вторым алгоритмом экстремальной группировки, хотя алгоритм и присоединяет обособленные параметры к другим группам. Наименее точные (среди трех сравниваемых алгоритмов) результаты дал метод корреляционных плеяд.

Исторически раньше возникшие различные варианты метода корреляционных плеяд являются в действительности несколько упрощенными эвристическими версиями более совершенных в математическом плане алгоритмов исследования структуры связей между компонентами многомерного признака, использующими графы-деревья и стохастические сети.

13.5. Построение сводного (интегрального) латентного показателя качества (или эффективности функционирования) сложной системы

13.5.1. Общая постановка задачи

И в профессиональной деятельности, и в своей повседневной жизни человек постоянно сталкивается с ситуациями, когда ему приходится сравнивать между собой и упорядочивать по некоторому *не поддающемуся непосредственному измерению* свойству ряд сложных систем. Речь может идти, в частности, о сравнении стран по уровню или качеству жизни, предприятий отрасли по эффективности их деятельности, сложных изделий (например, определенного программного или технического средства) — по обобщенной характеристике качества, специалистов — по эффективности их участия в выполнении поставленной задачи, участников игровых видов спорта — по уровню проявленного ими (в определенном состязании) мастерства и т. д. При этом общее представление о степени проявления анализируемого *латентного* (т. е. не поддающегося непосредственному измерению) свойства складывается как *результат определенного суммирования целого ряда частных (и поддающихся измерению) характеристик*, от которых зависит в конечном счете это свойство. Так, различные интегральные характеристики уровня благосостояния общества («уровень жизни», «качество жизни», «индекс человеческого развития») определяются набором показателей структуры и объемов потребления, качества населения и социальной сферы, экологического состояния окружающей среды. Точно так же эффективность работы предприятия определяется в основном совокупностью таких параметров, как себестоимость и реализуемость продукции, фондоотдача, текучесть кадров, количество рекламаций и т. п. В том же плане можно рассмотреть задачу измерения степени оптимальности социально-экономического поведения семьи в зависимости от значений параметров, определяющих структуру и объем семейного бюджета, времени и денег. Наконец, к типичным задачам тако-

го рода относится проблема измерения мастерства спортсмена в игровых видах спорта по значениям частных числовых характеристик его игровой деятельности (см. ниже пример с хоккеем).

Математико-статистической формализации подобных задач и вытекающим из нее рекомендациям по построению некоторого условного числового измерителя анализируемого латентного свойства сложной системы и посвящен данный пункт.

13.5.2. Сводный показатель («выходное качество») и его целевая функция

Пусть обобщенная сводная характеристика f анализируемого свойства объекта определяется набором частных критериев, задаваемых подпадающими учету и измерению переменными $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ (в дальнейшем будем называть их «входными»), однако сама эта характеристика является *латентной*, т.е. *не поддается непосредственному количественному измерению* (для нее не существует объективно обусловленной шкалы). Естественно предположить, что интуитивное экспертное (профессиональное) восприятие этой характеристики (обозначим его y) можно представить как несколько искаженное значение $f(x^{(1)}, \dots, x^{(p)})$, причем это искажение δ носит случайный характер и обусловлено как разрешающей способностью такого «измерительного прибора», каковым в данной схеме является эксперт, так и существованием ряда слабо влияющих на y , но не входящих в состав $X = (x^{(1)}, \dots, x^{(p)})$ «входных переменных». Тогда модель, связывающая между собой интуитивное представление о сводном показателе качества (y), сам сводный показатель ($f(X)$) как функцию от X и случайную погрешность $\delta(X)$, может быть определена в виде

$$y = f(X) + \delta(X). \quad (13.41)$$

Практически, не ограничивая общности данной схемы, можно принять естественные допущения относительно первых двух моментов остаточной случайной компоненты $\delta(X)$:

$$E \delta(X) \equiv 0, \quad D \delta(X) = \sigma^2(X) < \infty. \quad (13.41')$$

Тогда, очевидно, обобщенная (сводная) характеристика $f(X)$ может интерпретироваться как *регрессия y по X* , и если бы в качестве исходной статистической информации располагали бы наряду со значениями $X_i = (x_i^{(1)}, \dots, x_i^{(p)})^T$ и результатами регистрации соответствующих значений зависимой переменной y_i (i — номер наблюдения), то данная схема

непосредственно сводилась бы к обычной модели регрессии (см. гл. 10). Специфика модели (13.41), (13.41') состоит в том, что *вместо прямых измерений у можно получить (с помощью экспертов) лишь некоторые специального вида сведения о его значениях*, чаще всего — сравнительного плана (типа ранжировок или парных сравнений обследованных объектов по свойству y). Это обуславливает и более скромные претензии в отношении целей статистического анализа этой модели: вместо требуемого в регрессионном анализе восстановления (оценивания) функции $f(X)$ ставится *задача оценивания $f(X)$ с точностью до произвольного монотонного преобразования*.

О п р е д е л е н и е. *Целевой функцией* исследуемого обобщенного свойства («выходного качества») называется любое преобразование вида $\varphi(x^{(1)}, \dots, x^{(p)}) = \varphi(X)$, сохраняющее заданное соотношение порядка между анализируемыми объектами $O_1, O_2, \dots, O_n$ по усредненным значениям выходного качества, т.е. обладающее тем свойством, что из $f(X_{i_1}) \geq f(X_{i_2}) \geq \dots \geq f(X_{i_n})$ с необходимостью следует выполнение неравенств $\varphi(X_{i_1}) \geq \varphi(X_{i_2}) \geq \dots \geq \varphi(X_{i_n})$ и, наоборот, из последней серии неравенств вытекает выполнение соответствующих неравенств для $f(X_{i_k})$, $k = 1, 2, \dots, n$. Очевидно, данное здесь определение целевой функции неоднозначно. Действительно, если $\varphi(X)$ есть целевая функция и $U(\varphi)$ — любая взаимно-однозначная монотонно возрастающая функция, то всякая функция вида $\psi(X) = U[\varphi(X)]$ также будет целевой функцией. Это означает, что допущение о наличии определенной шкалы в измерении единого сводного показателя играет во многих случаях чисто вспомогательную роль и нацеливает на поиск, связанный с выявлением этой шкалы лишь с точностью до произвольного допустимого преобразования шкал. Ведь в соответствии с данным определением само значение целевой функции не отражает никакой реальной, физически содержательной количественной закономерности. Реальные закономерности отражаются только соотношениями «больше» или «меньше» между значениями этой функции для различных наборов величин входных параметров $X = (x^{(1)}, \dots, x^{(p)})^T$. Тем самым эти соотношения отражают предпочтение с точки зрения анализируемого выходного качества одних значений X перед другими. Поэтому в задачах, в которых возможно регулирование значений X (в некоторой допустимой области), наиболее рациональным управлением естественно признать то, которое максимизирует, при заданных ограничениях на X , значения целевой функции.

Данное определение целевой функции предполагает ее *содержательную* (экономическую, социально-экономическую, квалиметрическую, психологическую и т.д.) интерпретацию в зависимости от контекста иссле-

дования.

Итак, функция $f(X)$, с помощью которой можно было бы производить сравнительную оценку анализируемого «выходного качества» на рассматриваемых объектах, определена лишь с точностью до произвольного монотонного преобразования. Тем не менее для построения алгоритма ее восстановления было бы удобно параметризовать модель (13.41), т. е. определить параметрическое семейство $F = \{f(X; \Theta)\}$, в рамках которого будет производиться поиск целевой функции $f(X)$. Выбор этого параметрического семейства, как правило, не удастся подкрепить исчерпывающим теоретическим обоснованием, а потому с этого момента исследователь имеет дело не с целевой функцией $f(X)$, а с некоторой ее аппроксимацией $\hat{f}(X)$.

Имея в виду достаточную однородность обследуемых объектов по всем неучтенным переменным, т. е. по переменным, не вошедшим в состав $x^{(1)}, \dots, x^{(p)}$, и ограниченность интервала времени, в течение которого будем использовать искомую аппроксимацию целевой функции, а также реализуя идею разложения любой функции в ряд Тейлора, ограничимся в дальнейшем изложении аппроксимацией линейного вида, т. е.

$$\hat{f}(X; \Theta) = \theta_0 + \sum_{i=1}^p \theta_i x^{(i)}. \quad (13.42)$$

Коэффициенты $\Theta = (\theta_0, \theta_1, \dots, \theta_p)^T$ оцениваются статистически по исходным данным, структура и происхождение которых описываются ниже.

13.5.3. Исходные данные

Итак, пусть речь идет о построении непосредственно не поддающегося измерению единого сводного показателя эффективности функционирования (качества) объекта и пусть с этой целью были собраны исходные данные по n таким объектам: $O_1, O_2, \dots, O_n$. На основании этих исходных данных как раз и оцениваются параметры Θ искомой целевой функции $f(X; \Theta)$. Эти исходные данные состоят из двух частей: *экспертной* и *статистической* (отсюда название метода — экспертно-статистический).

Экспертная часть исходных данных. Эта часть исходных данных относится к сведениям о значениях случайной величины y_i (i — номер обследованного объекта) в модели (13.41) и получается с помощью специально организованного опроса экспертов и соответствующей статистической обработки экспертных оценок. При этом сведения об y_i ($i = 1, 2, \dots, n$) получают от экспертов в одной из следующих форм.

Форма (а) — наиболее информативный (а потому наиболее трудный для экспертов) вариант. Предусматривает получение *экспертных балльных оценок* выходного качества

$$\begin{aligned} & y_{11}, y_{21}, \dots, y_{n1}, \\ & y_{12}, y_{22}, \dots, y_{n2}, \\ & \dots\dots\dots, \\ & y_{1m}, y_{2m}, \dots, y_{nm}, \end{aligned} \quad (13.43a)$$

где y_{ij} — оценка выходного качества объекта O_i , полученная от j -го эксперта (здесь n — число оцениваемых объектов, m — число участвующих в оценке экспертов).

Форма (б) — средний по информативности (и по степени трудности для экспертов) вариант.

Предусматривает получение лишь *экспертных упорядочений* обследованных объектов по степени проявления в них анализируемого свойства, т. е. ранжировок вида

$$\begin{aligned} & R_{11}, R_{21}, \dots, R_{n1}, \\ & R_{12}, R_{22}, \dots, R_{n2}, \\ & \dots\dots\dots, \\ & R_{1m}, R_{2m}, \dots, R_{nm}, \end{aligned} \quad (13.43б)$$

где R_{ij} — ранг (место), присвоенный объекту O_i j -м экспертом в ряду из n обследованных объектов, упорядоченном этим экспертом по степени проявления анализируемого свойства.

Форма (в) — наименее информативный (и наименее трудный для экспертов) вариант. Информация от каждого (j -го) эксперта поступает в форме *булевой матрицы парных сравнений*

$$\gamma_j = (\gamma_{ik,j}), \quad i, k = \overline{1, n}, \quad j = 1, 2, \dots, m, \quad (13.43в)$$

где $\gamma_{ik,j}$ — результат парного сравнения j -м экспертом объектов O_i и O_k , может выражаться либо единицей, либо нулем по одному из следующих правил:

если эксперт производит сравнение объектов O_i и O_k типа их упорядочения по анализируемому свойству, то

$$\gamma_{ik,j} = \begin{cases} 1, & \text{если по мнению } j\text{-го эксперта } O_i \text{ не хуже } O_k, \\ 0 & \text{в противном случае;} \end{cases} \quad (13.43в')$$

если эксперт производит сравнение объектов O_i и O_k лишь с точки зрения принадлежности этих объектов к однородному (по анализируемому

свойству) классу, то

$$\gamma_{ik,j_s} = \begin{cases} 1, & \text{если объекты } O_i \text{ и } O_k \text{ однородны,} \\ 0 & \text{в противном случае.} \end{cases} \quad (13.43в'')$$

Вычислительные трудности, связанные с реализацией алгоритма оценивания параметров Θ искомой целевой функции $f(X; \Theta)$, естественно, возрастают по мере перехода от более информативных вариантов экспертной информации об y_i к менее информативным.

Статистическая часть исходных данных. Как выше уже отмечено, входные переменные (частные критерии) $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, на основании которых формируется представление об исследуемом выходном качестве, поддаются непосредственному измерению (регистрации) на каждом из обследуемых объектов. Поэтому, статистически обследовав анализируемые объекты $O_1, O_2, \dots, O_n$ по переменным $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, будем иметь *статистическую* часть исходных данных в виде матрицы (таблицы) типа «объект–свойство»:

$$\begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(p)} \\ \dots & \dots & \dots & \dots \\ x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(p)} \end{pmatrix}, \quad (13.44)$$

где $x_i^{(l)}$ — значение l -й входной переменной, зарегистрированное на i -м объекте.

Таким образом, приступая к оценке параметров Θ искомой целевой функции $f(X; \Theta)$ в модели (13.41), исследователь располагает исходной информацией об объектах $O_1, O_2, \dots, O_n$, состоящей из данных таблицы (13.44) и одного из вариантов (13.43а)–(13.43в) сведений об y_i .

13.5.4. Алгоритмические и вычислительные вопросы построения неизвестной целевой функции

Опишем вначале общую логическую схему оценивания параметров Θ целевой функции (13.42). Располагая конкретными значениями Θ_0 параметров Θ , для каждого объекта O_i можем вычислить величину среднего единого сводного показателя $f(X_i; \Theta_0)$ и далее, ориентируясь на сравнение значений $f(X_1; \Theta_0), f(X_2; \Theta_0), \dots, f(X_n; \Theta_0)$, получить основанную на целевой функции ранжировку объектов по искомому качеству

$$R_1(\Theta_0), R_2(\Theta_0), \dots, R_n(\Theta_0) \quad (13.45)$$

либо их разбиение на однородные (по f) классы, которое так же, как и уже имеющиеся экспертные разбиения (13.43в), может быть представлено в виде булевой матрицы $\gamma(\Theta_0; \Delta)$.

Оценку $\hat{\Theta}$ неизвестных параметров Θ предлагается подбирать таким образом, чтобы: 1) минимизировать расхождения в экспертных (y_{ij}) и полученных с помощью целевой функции ($f(X_i; \Theta)$) балльных оценках выходного качества (в варианте (а) экспертной информации); 2) максимизировать согласованность экспертных и полученной с помощью целевой функции ранжировок объектов по анализируемому выходному качеству (в варианте (б) экспертной информации); 3) минимизировать расхождения в экспертных и полученном с помощью целевой функции разбиениях объектов на классы (в варианте (в) экспертной информации).

Из сказанного следует, что экспертно-статистический метод построения единого сводного показателя нацелен на формализацию (в виде соответствующим образом подобранной целевой функции $f(X; \Theta)$) тех критерийных установок, которыми руководствовались привлеченные к контрольному эксперименту эксперты при формировании своих оценок. *Поэтому состоятельность и эффективность этого метода целиком зависят от компетентности и согласованности используемых в нем экспертных оценок.*

Оценивание неизвестных параметров целевой функции при балльных экспертных оценках выходного качества. В этом случае задача сводится к обычной схеме регрессионного анализа и соответственно к использованию метода наименьших квадратов. Действительно, располагая данными вида (13.43а)–(13.44), можем записать модель (13.41) в виде

$$\begin{cases} y_{ij} = f(X_i; \Theta) + \delta_j(X_i); \\ E \delta_j(X_i) \equiv 0, \quad D \delta_j(X_i) = \sigma_{ij}^2, \end{cases} \quad (13.46)$$

где величина (как правило, неизвестная) σ_{ij}^2 характеризует погрешность в оценке j -м экспертом выходного качества i -го объекта. Критерий метода наименьших квадратов дает нам оценку $\hat{\Theta}$ векторного параметра Θ как решение оптимизационной задачи вида (см. гл. 2 тома 2)

$$\sum_{j=1}^m \sum_{i=1}^n \frac{1}{\sigma_{ij}^2} (y_{ij} - f(X_i; \Theta))^2 \rightarrow \min_{\Theta}. \quad (13.47)$$

Если не располагают никакими сведениями относительно величин σ_{ij}^2 , то принимают упрощающее предположение $\sigma_{ij}^2 = \text{const}$, и задача (13.47) соответственно упрощается. В некоторых случаях удастся априори задаться «весами» ω_j , оценивающими сравнительную компетентность j -го

эксперта ($j = 1, 2, \dots, m$). Тогда эти веса вставляются в (13.47) в качестве сомножителей слагаемых вместо величин $1/\sigma_{ij}^2$.

Конкретные рекомендации по вычислительной реализации решения задач типа (13.47) приведены в гл. 2 тома 2.

Оценивание неизвестных параметров целевой функции при экспертных ранжировках и парных сравнениях объектов. Каждая экспертная ранжировка $R_{j\cdot} = (R_{1j}, R_{2j}, \dots, R_{nj})$ (j -я строка в (13.43в)) может быть представлена в виде булевой матрицы γ_j , в соответствии с правилом (13.43в). Поэтому в дальнейшем (в данном пункте) будем считать, что экспертная информация о выходном качестве объектов представлена в виде матриц парных сравнений вида (13.43в).

В общем случае задача состоит в том, чтобы на основе известных сравнений N пар объектов (не обязательно всех возможных пар из n объектов, т.е. N может быть меньше C_n^2) определить скалярную функцию $f(X; \Theta)$, такую, что парные сравнения, установленные по этой функции относительно тех же пар объектов, минимально (в смысле заданного критерия) отличались бы от экспертно установленных.

В случае парных сравнений в виде отношений предпочтения (см. правило (13.43в')) формирования элементов $\gamma_{ik,j}$, поставив на первое место в каждой из N экспертно оцененных пар лучший (не худший) объект, будем иметь пары (i_1, k_1) , значения целевых функций элементов которых должны были бы удовлетворять системе неравенств

$$\begin{cases} f(X_{i_q}; \Theta) - f(X_{k_q}; \Theta) \geq 0; \\ q = 1, 2, \dots, N. \end{cases} \quad (13.48)$$

Однако в общем случае эта система оказывается несовместной. Поэтому в каждое неравенство (i_q, k_q) вводится невязка

$$b_{i_q k_q}(\Theta) = \begin{cases} 0, & \text{если } f(X_{i_q}; \Theta) - f(X_{k_q}; \Theta) \geq 0; \\ -(f(X_{i_q}; \Theta) - f(X_{k_q}; \Theta)) & \text{в противном случае} \end{cases}$$

и вектор оценок $\hat{\Theta}$ определяется из условия минимума суммы невязок $\sum_{q=1}^N b_{i_q k_q}(\Theta)$ при некоторых ограничениях (типа нормировки) на компоненты искомого параметра Θ^1 .

¹ Алгоритм основан на результатах, изложенных в работе: Киселев Н. И. Экспертно-статистический метод определения функции предпочтения по результатам парных сравнений объектов // Алгоритмическое и программное обеспечение прикладного статистического анализа: Ученые записки по статистике. — М.: Наука, 1980. Т. 36. С. 111–122.

В случае парных сравнений, задающих разбиение объектов на однородные классы (см. правило (13.43в') формирования элементов γ_{ik,j_s}), матрицы $\gamma_1, \dots, \gamma_m$, задают m различных разбиений множества $\{O_1, O_2, \dots, O_n\}$ на классы, элементы каждого из которых близки по анализируемому выходному качеству. Для любых двух разбиений γ_s и γ_r может быть введена мера близости этих разбиений

$$d(\gamma_s, \gamma_r) = \frac{1}{2} \sum_{i,j=1}^n |\gamma_{ij,s} - \gamma_{ij,r}|.$$

Пусть $\hat{f}(X; \Theta) = \sum_{l=1}^p \theta_l x^{(l)}$ — некоторая линейная аппроксимация f . Задавшись некоторым $\varepsilon > 0$, можно с помощью $\hat{f}(X; \Theta)$ построить разбиение n объектов на классы. В один класс при этом попадут те объекты, у которых $0 \leq \hat{f}(X; \Theta) < \varepsilon$, в другой — те, у которых $\varepsilon \leq \hat{f}(X; \Theta) < 2\varepsilon$ и т. д. Полученное разбиение $\gamma(\varepsilon, \Theta)$ зависит, очевидно, от значений ε и Θ . Подбираются такие значения ε и Θ , чтобы величина $\sum_{j=1}^m d(\gamma_i, \gamma(\varepsilon; \Theta))$ была минимальна.

Для наилучшего выбора вектора коэффициентов Θ можно использовать также так называемый «метод голосования». При любом $\varepsilon > 0$ с помощью линейной функции $\hat{f}(X; \Theta) = \sum_{l=1}^p \theta_l x^{(l)}$ строится разбиение n объектов следующим образом. Пусть в разбиениях классы пронумерованы и $\gamma_j^{(\nu)}$ — ν -й класс в j -м экспертном разбиении. Для любого объекта X_i подсчитывается величина

$$\Gamma(X_i, \gamma_i^{(\nu)}) = \sum_{X_l \in \gamma_i^{(\nu)}} \tilde{\gamma}(X_i; X_l),$$

где

$$\tilde{\gamma}(X_i; X_l) = \begin{cases} 1, & \text{если } |\sum_{k=1}^p \theta_k (x_i^{(k)} - x_l^{(k)})| \leq \varepsilon; \\ 0, & \text{если } |\sum_{k=1}^p \theta_k (x_i^{(k)} - x_l^{(k)})| > \varepsilon. \end{cases}$$

Объект X_i относится к тому классу, для которого величина $\Gamma(X_i, \gamma_j^{(\nu)})$ максимальна. Полученное разбиение обозначим через $\gamma_j(\varepsilon; \Theta)$. Параметры ε и Θ подбираются из условия минимизации величины $\sum_{j=1}^m d(\gamma_j, \gamma_j(\varepsilon; \Theta))$ (при наличии априорных «весов компетентности» $v_1, \dots, v_m$ минимизируется взвешенная сумма $\sum_{j=1}^m v_j \cdot d(\gamma_j, \gamma_j(\varepsilon; \Theta))$). Используется алгоритм эвристического типа.

З а м е ч а н и е. Выше отмечалось, что успех описываемого подхода целиком зависит от качества экспертной части исходной информации. Поэтому прежде чем непосредственно приступить к процедурам оценивания параметров Θ целевой функции, необходимо тщательно исследовать

структуру и степень согласованности экспертных мнений. В варианте балльных оценок это сводится в основном к анализу резко выделяющихся наблюдений. В варианте ранжировок используется аппарат ранговой корреляции (см. гл. 11) в первую очередь для того, чтобы проверить гипотезу об отсутствии какой бы то ни было согласованности в упорядочениях различных экспертов. В варианте парных сравнений исследуется структура попарных расстояний между экспертными разбиениями на классы.

13.5.5. Примеры построения интегрального показателя с помощью экспертно-статистического метода

В п. 13.5.1 упоминались примеры реальных задач, в которых главной целью исследования было построение интегрального индикатора для некоторого сводного латентного «выходного качества» системы. Вернемся к некоторым из этих задач, чтобы дать краткое описание их решения или ссылку на литературный источник, в котором это решение подробно приводится.

1) **Построение интегрального показателя уровня мастерства хоккеиста, проявленного им в данном матче.** В рамках создания автоматизированной информационной системы для чемпионата мира по хоккею 1973 г. (системы «АИС-хоккей-73») рабочей группой Центрального экономико-математического института АН СССР была решена задача построения интегрального показателя уровня мастерства хоккеиста (отдельно для защитников и для нападающих) по одиннадцати частным показателям его игровой деятельности:

- $x^{(1)}$ — количество очков за результативность (по системе «гол+пас»);
- $x^{(2)}$ — число бросков в створ ворот противника;
- $x^{(3)}$ — число выигранных силовых единоборств;
- $x^{(4)}$ — число отборов шайбы у противника;
- $x^{(5)}$ — разность шайб («забито—пропущено») в микро матче данного игрока;
- $x^{(6)}$ — общее число точных передач;
- $x^{(7)}$ — число точных длинных первых передач;
- $x^{(8)}$ — число парированных бросков противника;
- $x^{(9)}$ — суммарное время участия хоккеиста в игре в период, когда его команда играла в большинстве или меньшинстве;
- $x^{(10)}$ — число удачно выполненных обводок;
- $x^{(11)}$ — сумма штрафного времени.

Экспертная часть исходной информации была получена в результате

заполнения экспертами специальных анкет, в которых уровень мастерства хоккеиста в данном матче оценивался по стобалльной системе (т. е. имелся вариант (13.43а) экспертной информации). «Съем» статистической части информации (13.44) по каждому из участников хоккейного матча производился специальными членами судейской комиссии.

В результате использования экспертно-статистического метода были получены следующие два варианта аппроксимаций для линейных целевых функций — *отдельно для оценки мастерства защитников* ($f_{\text{заш}}(X)$) *и нападающих* ($f_{\text{нап}}(X)$):

$$f_{\text{заш}}(X) = 10 + 4x^{(1)} + x^{(2)} + 4x^{(3)} + x^{(5)} + 0,2x^{(6)} + 3x^{(8)} + x^{(9)} + x^{(10)};$$

$$f_{\text{нап}}(X) = 8x^{(1)} + x^{(2)} + x^{(3)} + 0,5x^{(4)} + 0,2x^{(6)} + 4x^{(8)} + 2x^{(9)} + x^{(10)}.$$

Знание целевой функции позволяет в данном случае: 1) производить формализованную оценку мастерства хоккеиста, проявленного им в данном матче или в серии матчей, основанную только на знании отдельных числовых показателей, характеризующих его игру; 2) наиболее целесообразно строить индивидуальные планы тренировок, особое внимание уделяя совершенствованию тех компонентов игры, которые вошли в целевую функцию с относительно большими весами и за счет которых, следовательно, можно добиться наиболее существенного прироста в оценках мастерства.

Как и в любой работе такого профиля, в данной были последовательно реализованы следующие семь основных этапов:

- этап 1: постановка задачи;
- этап 2: предварительный отбор входных параметров;
- этап 3: организация экспертных обследований;
- этап 4: организация службы наблюдений, т. е. съема значений входных признаков;
- этап 5: вывод целевой функции (определение ее общего вида и вычисление весовых коэффициентов);
- этап 6: экспериментальная проверка адекватности целевой функции;
- этап 7: рабочая эксплуатация целевой функции.

Ход и результаты данной работы подробно описаны в [Айвазян С. А., Бежаева З. И., Староверов О. В., с. 218–223].

2) Построение сводного показателя эффективности деятельности промышленного предприятия. Объектами исследования являются 17 промышленных предприятий, специализирующихся на выпуске асинхронных двигателей переменного тока различного назначения. Цель исследования — построение единого сводного показателя экономической

эффективности работы предприятия в форме линейной функции от ряда частных показателей экономической эффективности. С помощью сочетания экспертного анализа и статистических методов снижения размерности (метода экстремальной группировки признаков п. 13.4.2, метода главных компонент, — п. 13.2) из априорного набора, состоящего из 22 частных показателей эффективности, было оставлено в качестве входных переменных экспертно-статистического метода восемь:

- $x^{(1)}$ — удельный вес продукции высшей категории качества в общем объеме товарной продукции (ТП) предприятия;
- $x^{(2)}$ — динамика¹ выпуска ТП на 1 рубль затрат;
- $x^{(3)}$ — выработка нормативно-чистой продукции (НЧП) на единицу промышленно-производственного персонала (ППП);
- $x^{(4)}$ — выполнение плана выпуска НЧП на единицу ППП;
- $x^{(5)}$ — динамика фондоотдачи;
- $x^{(6)}$ — выполнение плана выпуска ТП;
- $x^{(7)}$ — выполнение плана по оборачиваемости нормируемых оборотных средств (отношение фактического числа оборотов к нормативному);
- $x^{(8)}$ — выполнение плана по балансовой прибыли.

Из двенадцати привлеченных к задаче экспертов пять дали оценку интегральной эффективности деятельности анализируемых предприятий (по результатам их работы в 1982 и 1983 гг.) в десятибалльной системе, остальные семь проранжировали предприятия, причем четверо из них дополнительно представили результаты своих парных сравнений. Поэтому реализовывались все три версии оценивания неизвестных коэффициентов линейной целевой функции. Приведем здесь для примера один из полученных с помощью ЭСМ вариантов решения, а именно вариант, ориентированный на экспертные балльные оценки только одного из экспертов: $f(X; \Theta) = -2,94 + 0,10x^{(1)} + 0,07x^{(2)} + 0,55x^{(3)} - 0,13x^{(4)} + 0,06x^{(5)} - 0,03x^{(6)} - 0,07x^{(7)} - 0,02x^{(8)}$.

Мера согласованности балльного оценивания предприятий, произведенного с помощью этого эксперта и данной целевой функции, характеризуется величиной коэффициента корреляции, равной 0,77. Обращает на себя внимание тот факт, что формализация критерийных установок эксперта показала практическое игнорирование им при формировании интегральной оценки эффективности функционирования предприятия всех

¹ Показатель динамики исчисляется как отношение прироста анализируемого показателя в данном периоде к его величине в предшествующем.

частных показателей выполнения плана (т.е. входных переменных $x^{(4)}$, $x^{(6)}$, $x^{(8)}$). В то же время если бы этому специалисту предложили непосредственно (экспертно) оценить значимость коэффициентов при этих переменных, то результат, можно не сомневаться, был бы совсем иным!

13.6. Многомерное шкалирование

13.6.1. Постановка задачи метрического многомерного шкалирования

Пусть исходная информация об объектах задана в форме *матрицы их попарных сравнений* (9.2), где элемент γ_{ij} этой матрицы интерпретируется как евклидово расстояние между объектом O_i и объектом O_j ($i, j = 1, 2, \dots, n$), т.е.

$$\gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \dots & \gamma_{1n} \\ \gamma_{21} & \gamma_{22} & \dots & \gamma_{2n} \\ \dots & \dots & \dots & \dots \\ \gamma_{n1} & \gamma_{n2} & \dots & \gamma_{nn} \end{pmatrix}, \quad (13.49)$$

где

$$\gamma_{ij} = \sqrt{\sum_{k=1}^p (x_i^{(k)} - x_j^{(k)})^2}. \quad (13.50)$$

Однако ни координаты объектов $X_i = (x_i^{(1)}, \dots, x_i^{(p)})^T$, ни даже размерность их признакового пространства p нам не известны. Требуется на основании известных данных (13.49) восстановить неизвестную размерность p анализируемого признакового пространства и приписать каждому объекту O_i координаты $X_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)})^T$ таким образом, чтобы вычисленные по формуле (13.50) попарные евклидовы расстояния по возможности совпали бы с заданными матрицей γ . При этом речь идет о восстановлении координат $X_1, X_2, \dots, X_n$ с *точностью до ортогонального преобразования*, т.к. при ортогональном преобразовании системы координат попарные расстояния между объектами не меняются. Таким образом, *цель методов метрического многомерного шкалирования состоит в том, чтобы отобразить информацию о конфигурации исходных многомерных данных, заданную матрицей расстояний γ , в виде геометрической конфигурации n точек в соответствующем многомерном пространстве.*

13.6.2. Решение задачи метрического многомерного шкалирования

Введем в рассмотрение центрированную матрицу искомым наблюдений

$$X_{\Pi} = \begin{pmatrix} x_1^{(1)} - \bar{x}^{(1)} & x_1^{(2)} - \bar{x}^{(2)} & \dots & x_1^{(p)} - \bar{x}^{(p)} \\ x_2^{(1)} - \bar{x}^{(1)} & x_2^{(2)} - \bar{x}^{(2)} & \dots & x_2^{(p)} - \bar{x}^{(p)} \\ \dots & \dots & \dots & \dots \\ x_n^{(1)} - \bar{x}^{(1)} & x_n^{(2)} - \bar{x}^{(2)} & \dots & x_n^{(p)} - \bar{x}^{(p)} \end{pmatrix}, \quad (13.51)$$

а также матрицу $B = (b_{ij})$, $i, j = 1, 2, \dots, n$, скалярных произведений

$$b_{ij} = (X_i - \bar{X})^T (X_j - \bar{X}) = \sum_{k=1}^p (x_i^{(k)} - \bar{x}^{(k)})(x_j^{(k)} - \bar{x}^{(k)}). \quad (13.52)$$

Метод определения искомым координат $X_1, X_2, \dots, X_n$ анализируемых объектов (с точностью до ортогонального преобразования), а заодно и размерности p пространства, в которое они отображаются, основан не на непосредственном использовании матрицы исходных данных γ (см. (13.49)), а на преобразовании ее в матрицу B . Можно показать, что элементы матриц γ и B связаны между собой соотношением

$$b_{ij} = \frac{1}{2} \left(-\gamma_{ij}^2 + \frac{1}{n} \sum_{i=1}^n \gamma_{ij}^2 + \frac{1}{n} \sum_{j=1}^n \gamma_{ij}^2 - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \gamma_{ij}^2 \right). \quad (13.53)$$

Далее используются следующие свойства матрицы B :

- (i) матрица B неотрицательно определена;
- (ii) ранг матрицы B равен размерности p искомого пространства отображения;
- (iii) ненулевые собственные числа $\lambda_1, \lambda_2, \dots, \lambda_p$ матрицы B , упорядоченные в порядке убывания, совпадают с соответствующими собственными числами матрицы $S^2 = X_{\Pi}^T X_{\Pi}$, где матрица X_{Π} определена соотношением (13.51); отметим, что матрица $\hat{\Sigma}_X = S^2/n$ есть выборочная матрица ковариаций искомого вектора признаков X ;
- (iv) пусть $l_r = (l_{r1}, l_{r2}, \dots, l_{rp})^T$ есть r -й собственный вектор матрицы S^2 , соответствующий r -му по величине собственному значению λ_r ($r = 1, 2, \dots, p$); тогда вектор $Z^{(r)} = (z_1^{(r)}, z_2^{(r)}, \dots, z_n^{(r)})^T$ значений r -й главной компоненты вектора X будет

$$Z^{(r)} = X_{\Pi} l_r; \quad (13.54)$$

при этом, если $l_r^B = (l_{r1}^B, l_{r2}^B, \dots, l_{rn}^B)^T$ — r -й собственный вектор матрицы B , соответствующий тому же собственному числу λ_r (т. е. $B l_r^B = \lambda_r l_r^B$), то

$$Z^{(r)} = \sqrt{\lambda_r} l_r^B, \quad r = 1, 2, \dots, p. \quad (13.55)$$

В конечном счете из приведенных свойств следует, что, решая проблему собственных чисел и собственных векторов для матрицы B и ограничиваясь ненулевыми собственными числами $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$, получим координатное представление (13.55) анализируемых объектов в пространстве главных компонент искомого многомерного признака X . А поскольку мы можем решить задачу *только с точностью до ортогонального преобразования*, то представление объектов в виде (13.55) и будет решением проблемы метрического многомерного шкалирования. Заметим, что из построения следует возможность следующего выражения элементов b_{ij} через $z_i^{(r)}$ и $z_j^{(r)}$ ($r = 1, 2, \dots, p$):

$$b_{ij} = \sum_{r=1}^p z_i^{(r)} z_j^{(r)}. \quad (13.56)$$

Если же мы хотим представить имеющиеся данные в пространстве заданной (но меньшей, чем p) размерности p' , то можно показать, что, беря только первые p' координат в выражении (13.55), мы обеспечим максимальное (среди прочих представлений исходных статистических данных в пространстве размерности $p' < p$) приближение их геометрической структуры, заданной элементами b_{ij} матрицы B в смысле критерия

$$\sum_{i=1}^n \sum_{j=1}^n (b_{ij} - b_{ij}^{(p')}(X))^2. \quad (13.57)$$

В соотношении (13.57) величины $b_{ij}^{(p')}(X)$ означают значения геометрических характеристик, вычисленных по формулам (13.53), в которых стоящие в правых частях попарные расстояния γ_{ij} вычислены в координатах предлагаемого p' -мерного пространства (как следует из (13.56), в случае $p' = p$ этот критерий принимает нулевое значение при использовании представления (13.55)).

13.6.3. Понятие о неметрическом многомерном шкалировании (МШ)

В неметрическом МШ предполагается, что различия (близости) γ_{ij} измерены в ординальной шкале, так что важен только ранговый порядок различий, а не сами их численные значения. Процедуры неметрического МШ стремятся построить такую геометрическую конфигурацию точек в пространстве заданной размерности p' , чтобы ранговый порядок попарных расстояний между ними, по возможности, минимально отличался от того порядка, который задан матрицей γ . Одна процедура неметрического МШ отличается от другой выбором вида критерия различия двух разных упорядочений.

ВЫВОДЫ

1. В исследовательской и практической статистической деятельности часто приходится иметь дело с исходными данными *высокой размерности*, т.е. с ситуациями, когда число регистрируемых на *каждом* из статистически обследованных объектов показателей составляет несколько десятков, а иногда — сотни и даже тысячи. В подобных ситуациях легко объяснимо желание исследователя *существенно снизить размерность анализируемого признакового пространства*, т.е. перейти от исходного набора показателей к небольшому числу *вспомогательных переменных* (которые либо отбираются из числа исходных, либо строятся по определенному правилу по совокупности исходных показателей), по которым впоследствии он мог бы достаточно точно воспроизвести интересующие его свойства анализируемого массива данных. Одним из наиболее распространенных методов снижения размерности исследуемого признакового пространства является *метод главных компонент*.

2. Имеется по меньшей мере три основных типа принципиальных предпосылок, обуславливающих возможность практически «безболезненного» перехода от большого числа исходных показателей состояния (поведения, качества, эффективности функционирования) анализируемого объекта к существенно меньшему числу наиболее информативных переменных. Это, во-первых, *дублирование информации, доставляемой сильно взаимосвязанными показателями*; во-вторых, *неинформативность показателей, мало меняющихся при переходе от одного объекта к другому (малая вариабельность показателя)*; в-третьих, *возможность агрегирования*, т.е. простого или взвешенного суммирования некоторых физически однотипных показателей.

3. В оптимизационной постановке задачи снижения размерности

решение, получаемое с помощью метода главных компонент, максимизирует критерий информативности, определяемый суммарной дисперсией заданного (небольшого) числа искоемых вспомогательных переменных (при соответствующих условиях их нормировки). Для вычисления k -й главной компоненты $z^{(k)}(X)$ ($k = 1, \dots, p$) следует найти собственный вектор $l_k = (l_{k1}, \dots, l_{kp})$ ковариационной матрицы Σ исходного набора показателей $X = (x^{(1)}, \dots, x^{(p)})^\top$, т.е. решить систему уравнений $(\Sigma - \lambda_k I)l_k^\top = 0$, где λ_k — k -й по величине корень (при их расположении в порядке убывания) характеристического уравнения $|\Sigma - \lambda I| = 0$. Компоненты l_{kj} ($j = \overline{1, p}$) собственного вектора l_k являются искомыми весовыми коэффициентами, с помощью которых осуществляется переход от исходных показателей $x^{(1)}, \dots, x^{(p)}$ к главной компоненте $z^{(k)}(X)$, т.е. $z^{(k)}(X) = l_k \cdot X$.

4. Основные числовые характеристики вектора $Z = (z^{(1)}, \dots, z^{(p)})^\top$ главных компонент могут быть выражены через основные числовые характеристики исходных показателей и собственные числа их ковариационной матрицы Σ (см. п. 13.2.3).

5. Вектор p' ($p' < p$) первых главных компонент $Z^{(p')}(X) = (z^{(1)}(X), \dots, z^{(p')}(X))^\top$ обладает рядом оптимальных свойств, среди которых отметим следующие:

а) *свойство наименьшей ошибки автопрогноза или наилучшей самовоспроизводимости*: с помощью p' первых главных компонент $z^{(1)}, \dots, z^{(p')}$ исходных показателей $x^{(1)}, \dots, x^{(p)}$ ($p' < p$) достигается наилучший (в определенном смысле) прогноз этих показателей среди всех прогнозов, которые можно построить с помощью p' линейных комбинаций набора из p произвольных признаков;

б) *свойство наименьшего искажения некоторых геометрических характеристик совокупности исходных многомерных наблюдений* $X_1, \dots, X_n$ при их проецировании в пространство меньшей размерности, натянутое на p' первых главных компонент $z^{(1)}, \dots, z^{(p')}$.

6. Главные компоненты, построенные не по истинной ковариационной матрице Σ вектора исходных показателей $X = (x^{(1)}, \dots, x^{(p)})^\top$, а по ее выборочному аналогу (оценке) $\hat{\Sigma}$, называются *выборочными главными компонентами* и в определенных (достаточно широких) условиях обладают (вместе с собственными числами и векторами матрицы $\hat{\Sigma}$) всеми традиционными свойствами «хороших» оценок: состоятельностью, асимптотической эффективностью, асимптотической нормальностью.

7. Геометрически определение первой главной компоненты равносиль-

но построению новой координатной оси $Oz^{(1)}$ таким образом, чтобы она шла в направлении наибольшего разброса исходных данных, т.е. в направлении вытянутости анализируемого «облака» многомерных наблюдений. Затем среди направлений, перпендикулярных к $Oz^{(1)}$, отыскивается направление «наибольшей вытянутости» $Oz^{(2)}$ и т.д. Если характер вытянутости анализируемого «облака» данных в исходном признаковом пространстве существенно отличен от линейного, то *линейная* модель главных компонент может оказаться неэффективной.

8. Главные компоненты используются при решении следующих основных типов задач анализа данных:

1) *упрощение, сокращение размерностей анализируемых моделей* статистического исследования зависимостей или классификации с целью облегчения счета и интерпретации получаемых статистических выводов;

2) *наглядное представление (визуализация)* исходных многомерных данных, получаемое с помощью их проецирования в пространство, натянутое на первую, первые две или первые три главные компоненты;

3) *предварительная ортогонализация* объясняющих переменных в задачах построения регрессионных зависимостей как средство «борьбы» с мультиколлинеарностью (см. гл. 2 в томе 2);

4) *сжатие объемов хранимой статистической информации*.

9. Различные версии моделей и методов факторного анализа (центроидный, максимального правдоподобия, экстремальной группировки параметров, корреляционных плеяд и др.) основаны на общей базовой идее, в соответствии с которой значения всех признаков $x^{(1)}, \dots, x^{(p)}$ анализируемого набора формируются под воздействием сравнительно небольшого числа одних и тех же (общих) факторов $f^{(1)}, \dots, f^{(p')}$, не поддающихся, правда, непосредственному измерению (и потому называющихся латентными). В определенном смысле эти общие факторы выступают в роли причин, а наблюдаемые (анализируемые) признаки — в роли следствий.

10. Поскольку число общих (латентных) факторов существенно меньше числа анализируемых признаков, то методы факторного анализа в конечном счете нацелены (так же как и метод главных компонент) на *снижение размерности анализируемого признакового пространства*.

11. Статистическая реализация модели факторного анализа предусматривает последовательное решение вопросов *существования* такой модели, ее *идентификации* (т.е. возможности ее однозначного восстановления по исходным статистическим данным), *алгоритмического определения ее структурных параметров* (т.е. определения способа вычисления неизвестных параметров модели при точно известной ковариационной матрице анализируемого многомерного признака) и *из статистической*

оценки по имеющимся наблюдениям, включая статистические оценки для самих общих (латентных) факторов.

12. Наиболее распространенной в практике статистических исследований и наиболее теоретически разработанной является *каноническая модель* факторного анализа, в которой признаки линейно зависят от факторов, факторы взаимно некоррелированы между собой и со случайными остатками модели, а случайные остатки в свою очередь взаимно некоррелированы и нормально распределены.

13. Между методом главных компонент и линейной моделью факторного анализа имеется идейная общность: и тот и другой метод можно рассматривать как метод аппроксимации набора анализируемых перемещенных переменных с помощью линейных функций от сравнительно небольшого числа одних и тех же вспомогательных переменных (главных компонент — в одном методе и общих факторов — в другом). Их различие — лишь в конкретизации критерия точности аппроксимации.

14. Наиболее «узкие места» в практической дееспособности модели факторного анализа связаны с решением задачи оценки числа p' общих факторов модели и с содержательной интерпретацией найденных общих факторов. Для успешного решения последней задачи широко пользуются неоднозначностью (с точностью до ортогонального преобразования) определения общих факторов и соответственно возможностью их разнообразных «вращений» в факторном пространстве.

15. Наряду с математико-статистическими методами снижения размерности, т. е. с методами, допускающими описание и интерпретацию в терминах строгой вероятностной модели, существуют и широко используются в статистической практике так называемые *эвристические методы*. Свое название они оправдывают тем, что порождаются обычно некоторыми частными целевыми установками, выраженными в виде *установленных на содержательно-субъективном уровне* оптимизируемых критериев качества решения задачи. К таким методам, в частности, относятся методы экстремальной группировки, метод корреляционных плеяд.

16. Задача построения не поддающегося непосредственному измерению *интегрального* (агрегатного) сводного показателя y эффективности функционирования (качества) объекта по заданным значениям частных критериальных характеристик $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ анализируемого свойства может рассматриваться как задача снижения размерности исследуемого признакового пространства *до единицы*. Эта же задача может быть сформулирована в терминах построения целевой функции анализируемого обобщенного свойства исследуемых объектов.

17. Базовая идея *экспертно-статистического метода* построения

единого сводного показателя эффективности функционирования (качества) объекта заключается в «настройке» искомых коэффициентов θ_j ; целевой функции $\varphi(X)$ на заданную (в различной форме) экспертную информацию, относящуюся к сравнению статистически обследованных объектов по анализируемому интегральному свойству. Название метода объясняется тем, что его реализация основана как на *статистической информации* об объектах, так и на *экспертной* (это представленные в той или иной форме экспертные оценки анализируемого интегрального свойства y).

18. Вычислительная реализация экспертно-статистического метода (т.е. алгоритм определения искомых «весов» θ_j) сводится к известному *методу наименьших квадратов* лишь в тех сравнительно редких случаях, когда от экспертов удастся получить *балльные оценки* $y_1, \dots, y_n$, анализируемого интегрального свойства *по каждому из исследуемых объектов*. Если же в распоряжении исследователя лишь *сравнительные оценки* объектов по анализируемому свойству (упорядочения, парные сравнения, классификации), то вычислительная процедура по определению коэффициентов θ_j существенно усложняется (ее описание и обоснование требуют специальной разработки).

19. Экспертно-статистический метод имеет широкий диапазон возможных применений, однако необходимым условием его достоверности и эффективности является *четкое определение анализируемого интегрального свойства и компетентность используемых экспертных мнений*.

20. *Многомерное шкалирование*, — совокупность методов, позволяющих по заданной информации о мерах различия (близости) между объектами рассматриваемой совокупности приписывать каждому из этих объектов вектор характеризующих его количественных показателей; при этом размерность искомого координатного пространства задается заранее, а «погружение» в него анализируемых объектов производится таким образом, чтобы структура взаимных различий (близостей) между ними, измеренных с помощью приписываемых им вспомогательных координат, в среднем наименее отличалась бы от заданной в смысле того или иного функционала качества. Процедуры многомерного шкалирования применяются, когда данные заданы в виде матрицы попарных расстояний между объектами или удаленностей или их порядковых отношений. В первом случае используются методы так называемого метрического шкалирования, а во втором — неметрического шкалирования.

Приложения

**Приложение 1. Таблицы математической
статистики**

**Приложение 2. Необходимые сведения
из матричной алгебры**

ПРИЛОЖЕНИЕ 1. ТАБЛИЦЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Таблица П1.1. Значения функции плотности $\phi(x) = \phi(x; 0; 1)$ стандартного нормального закона распределения. $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

x	$\phi(x)$	x	$\phi(x)$	x	$\phi(x)$	x	$\phi(x)$	x	$\phi(x)$	x	$\phi(x)$
0,00	0,3989	0,50	0,3521	1,00	0,2420	1,50	0,1295	2,00	0,0540	2,50	0,0175
0,05	0,3984	0,55	0,3429	1,05	0,2299	1,55	0,1200	2,05	0,0488	2,55	0,0154
0,10	0,3970	0,60	0,3332	1,10	0,2179	1,60	0,1109	2,10	0,0440	2,60	0,0136
0,15	0,3945	0,65	0,3230	1,15	0,2059	1,65	0,1023	2,15	0,0396	2,65	0,0119
0,20	0,3910	0,70	0,3123	1,20	0,1942	1,70	0,0940	2,20	0,0355	2,70	0,0104
0,25	0,3867	0,75	0,3011	1,25	0,1826	1,75	0,0863	2,25	0,0317	2,75	0,0091
0,30	0,3814	0,80	0,2897	1,30	0,1714	1,80	0,0790	2,30	0,0283	2,80	0,0079
0,35	0,3752	0,85	0,2780	1,35	0,1604	1,85	0,0721	2,35	0,0252	2,85	0,0069
0,40	0,3683	0,90	0,2661	1,40	0,1497	1,90	0,0656	2,40	0,0224	2,90	0,0060
0,45	0,3605	0,95	0,2541	1,45	0,1394	1,95	0,0596	2,45	0,0198	2,95	0,0051
										3,00	0,0044

З а м е ч а н и е 1. Значения функции плотности $\phi(x_0; a; \sigma^2)$ нормального закона со средним a и дисперсией σ^2 подсчитывается по формуле

$$\phi(x_0; a; \sigma^2) = \frac{1}{\sigma} \phi\left(\frac{x_0 - a}{\sigma}; 0; 1\right) = \frac{1}{\sigma} \phi\left(\frac{x_0 - a}{\sigma}\right) \quad (\text{П1.1})$$

(если величина аргумента $(x_0 - a)/\sigma$ попадает *между* табличными значениями x , то для определения $\phi(\frac{x_0 - a}{\sigma})$ пользуются линейной интерполяцией функции $\phi(x)$).

З а м е ч а н и е 2. При определении значений функции $\phi(x)$ для отрицательных величин аргумента x следует использовать тождество (выражающее свойство *четности* функции $\phi(x)$)

$$\phi(-x) \equiv \phi(x). \quad (\text{П1.2})$$

П р и м е р П1.1. Требуется определить значение $\phi(x_0; a; \sigma^2)$ при $x_0 = 3,36$, $a = 1$ и $\sigma^2 = 4$.

Решение: $x = \frac{x_0 - a}{\sigma} = \frac{3,36 - 1}{2} = 1,18$. Два окаймляющих x соседних табличных значения аргумента — это $x_1 = 1,15$ и $x_2 = 1,20$, поэтому, используя линейную интерполяцию, получаем:

$$\phi(1,18) = \phi(1,15) - \frac{1,18 - 1,15}{1,20 - 1,15} [\phi(1,15) - \phi(1,20)] = 0,199.$$

Значение $\phi(3,36; 1; 4)$ получаем по формуле (П1.1):

$$\phi(x_0 = 3,36; 1; 4) = \frac{1}{2} \phi(1,18) = \frac{1}{2} 0,199 = 0,0995.$$

Таблица П1.2. Значения функции $\Phi(x) = \Phi(x; 0; 1)$ стандартного нормального распределения.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt.$$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,500000	1,00	0,841345	2,00	0,977250
0,05	0,519939	1,05	0,853141	2,05	0,979818
0,10	0,539828	1,10	0,864334	2,10	0,982136
0,15	0,559618	1,15	0,874928	2,15	0,984222
0,20	0,579260	1,20	0,884930	2,20	0,986097
0,25	0,589706	1,25	0,894350	2,25	0,987776
0,30	0,617911	1,30	0,903200	2,30	0,989276
0,35	0,636831	1,35	0,911492	2,35	0,990613
0,40	0,655422	1,40	0,919243	2,40	0,991802
0,45	0,673645	1,45	0,926471	2,45	0,992857
0,50	0,691463	1,50	0,933193	2,50	0,993790
0,55	0,708840	1,55	0,939429	2,55	0,994614
0,60	0,725747	1,60	0,945201	2,60	0,995339
0,65	0,742154	1,65	0,950528	2,65	0,995975
0,70	0,758036	1,70	0,955434	2,70	0,996533
0,75	0,773373	1,75	0,959941	2,75	0,997020
0,80	0,788145	1,80	0,964070	2,80	0,997445
0,85	0,802338	1,85	0,967843	2,85	0,997814
0,90	0,815940	1,90	0,971283	2,90	0,998134
0,95	0,828944	1,95	0,974412	2,95	0,998411
				3,00	0,998650

З а м е ч а н и е 1. Значение функции распределения $\Phi(x_0; a; \sigma^2)$ нормального закона при среднем значении a и дисперсии σ^2 в заданной точке x_0 подсчитывается по значениям функции $\Phi(x) = \Phi(x; 0; 1)$ с помощью формулы

$$\Phi(x_0; a; \sigma^2) = \Phi\left(\frac{x_0 - a}{\sigma}; 0; 1\right) = \Phi\left(\frac{x_0 - a}{\sigma}\right) \quad (\text{П1.3})$$

(если величина аргумента $(x_0 - a)/\sigma$ попадает между табличными значениями x , то для определения $\Phi((x_0 - a)/\sigma)$ пользуются линейной интерполяцией функции $\Phi(x)$).

З а м е ч а н и е 2. При определении значений функции $\Phi(x)$ для отрицательных величин аргумента x следует использовать тождество

$$\Phi(x) \equiv 1 - \Phi(-x). \quad (\text{П1.4})$$

П р и м е р П1.2. В условиях примера П1.1 имеем:

$$\begin{aligned} \Phi(x_0 = 3,36; a = 1; \sigma^2 = 4) &= \Phi(1,18) \\ &= \Phi(1,15) + \frac{1,18 - 1,15}{1,20 - 1,15} [\Phi(1,20) - \Phi(1,15)] = 0,881. \end{aligned}$$

Таблица П1.3. Значения q -квантилей u_q стандартного нормального распределения.

q	u_q	q	u_q	q	u_q	q	u_q
0,50	0,000000	0,70	0,524401	0,90	1,281552	0,983	120072
51	025069	71	553385	91	340755	984	144411
52	050154	72	582842	92	405072	0,985	2,170090
53	075270	73	612813	93	475791	986	197286
54	100434	74	643345	94	554774	987	226212
0,55	0,125661	0,75	0,674490	0,95	1,644854	988	257129
56	150969	76	706303	96	750686	989	290368
57	176374	77	738847	97	880794	0,990	2,326348
58	201893	78	772193	971	895698	991	365618
59	227545	79	806421	972	911036	992	408916
0,60	0,253347	0,80	0,841621	973	926837	993	457263
61	279319	81	877896	974	943134	994	512144
62	305481	82	915365	0,975	1,959964	0,995	2,575829
63	331853	83	954165	976	977368	996	652070
64	358459	84	994458	977	995393	997	747781
0,65	0,385320	0,85	1,036433	978	2,014091	998	878162
66	412463	86	080319	979	033520	999	3,090232
67	439913	87	126391	0,980	2,053749		
68	467699	88	174987	981	074855		
69	495850	89	226528	982	096927		

Пример П1.3. Найти 0,9-квантиль $u_{0,9}$. Величину $u_{0,9}$ находим из таблицы в графе, расположенной справа от соответствующего значения $q = 0,9$, т. е. $u_{0,9} = 1,281552$.

З а м е ч а н и е 1. Если заданная величина q попадает между двумя соседними табличными значениями q_1 и q_2 ($q_1 < q_2$; это может случиться при графической проверке нормальности распределения), то следует воспользоваться линейной интерполяцией, а именно формулой

$$u_q = u_{q_1} + \frac{q - q_1}{q_2 - q_1} (u_{q_2} - u_{q_1}).$$

З а м е ч а н и е 2. При нахождении q -квантилей для значений $q < 0,5$ следует воспользоваться соотношением $u_q = -u_{1-q}$. Например, $u_{0,4} = -u_{1-0,4} = -u_{0,6} = -0,25335$.

З а м е ч а н и е 3. При отыскании 100 Q %-ных точек w_Q следует воспользоваться соотношением $w_Q = u_{1-Q}$. Например, $w_{0,05} = u_{0,95} = 1,64485$.

Таблица П.1.4. Значения 100Q%-ных точек $\chi^2_Q(\nu)$ χ^2 -распределения с ν степенями свободы.

ν	Q									
	0,995	0,990	9,975	0,950	0,900	0,100	0,050	0,025	0,010	0,005
1	392704 · 10 ⁻¹⁰	157088 · 10 ⁻⁹	982069 · 10 ⁻⁹	393214 · 10 ⁻⁸	0,0157908	2,70554	3,84146	5,02389	6,63490	7,87944
2	0,0100251	0,0201007	0,0506356	0,102587	0,210720	4,60517	5,99147	7,37776	9,21034	10,5966
3	0,0717212	0,114832	0,215795	0,351846	0,584375	6,25139	7,81473	9,34840	11,3449	12,8381
4	0,206990	0,297110	0,484419	0,710721	1,063623	7,77944	9,48773	11,1433	13,2767	14,8602
5	0,411740	0,554300	0,831211	1,145476	1,61031	9,23635	11,0705	12,8325	15,0863	16,7496
6	0,675727	0,872085	1,237347	1,63539	2,20413	10,6446	12,5916	14,4494	16,8119	18,5476
7	0,989265	1,239043	1,68987	2,16735	2,83311	12,0170	14,0571	16,0128	18,4753	20,2777
8	1,344419	1,646482	2,17973	2,73264	3,48954	13,3616	15,5073	17,5346	20,0902	21,9550
9	1,734926	2,087912	2,70039	3,32511	4,16816	14,6837	16,9190	19,0228	21,6660	23,5893
10	2,15585	2,55821	3,24697	3,94030	4,86518	15,9871	18,3070	20,4831	23,2093	25,1882
11	2,60321	3,05347	3,81575	4,57481	5,57779	17,2750	19,6751	21,9200	24,7250	26,7569
12	3,07382	3,57056	4,40379	5,22603	6,30380	18,5494	21,0261	23,3357	26,2170	28,2995
13	3,56503	4,10691	5,00874	5,89186	7,04150	19,8119	22,3621	24,7356	27,6883	29,8194
14	4,07468	4,66043	5,62872	6,57063	7,78953	21,0642	23,6848	26,1190	29,1413	31,3193
15	4,60094	5,22935	6,26214	7,26094	8,54675	22,3072	24,9958	27,4884	30,5779	32,8013
16	5,14224	5,81221	6,90766	7,96164	9,31223	23,5418	26,2962	28,8454	31,9999	34,2672
17	5,69724	6,40776	7,56418	8,67176	10,0852	24,7690	27,5871	30,1910	33,4087	35,7185
18	6,26481	7,01491	8,23075	9,39046	10,8649	25,9894	28,8693	31,5264	34,8053	37,1554
19	6,84398	7,63273	8,90655	10,1170	11,6509	27,2035	30,1435	32,8523	36,1908	38,5822

20	7, 43386	8, 26040	9, 59083	10, 8508	12, 4426	28, 4120	31, 4104	34, 1696	37, 5662	39, 9968
21	8, 03366	8, 89720	10, 28293	11, 5913	13, 2396	29, 6151	32, 6705	35, 4789	36, 9321	41, 4010
22	8, 64272	9, 54249	10, 9823	12, 3380	14, 0415	30, 8133	33, 9244	36, 7807	40, 2894	42, 7956
23	9, 26042	10, 19567	11, 6885	13, 0905	14, 8479	32, 0069	35, 1725	38, 0757	41, 6384	44, 1813
24	9, 88623	10, 8564	12, 4011	13, 8484	15, 6587	33, 1963	36, 4151	39, 3641	42, 9798	45, 5585
25	10, 5197	11, 5240	13, 1197	14, 6114	16, 4734	34, 3816	37, 6525	40, 6465	44, 3141	46, 9278
26	11, 1603	12, 1981	13, 8439	15, 3791	17, 2919	35, 5631	38, 8852	41, 9232	45, 6417	48, 2899
27	11, 8076	12, 8786	14, 5733	16, 1513	18, 1138	36, 7412	40, 1133	43, 1944	46, 9630	49, 6449
28	12, 4613	13, 5648	15, 3079	16, 9279	18, 9392	37, 9159	41, 3372	44, 4607	48, 2782	50, 9933
29	13, 1211	14, 2565	16, 0471	17, 7083	19, 7677	39, 0875	42, 5569	45, 7222	49, 5879	52, 3356
30	13, 7867	14, 9535	16, 7908	18, 4926	20, 5992	40, 2560	43, 7729	46, 9792	50, 8922	53, 6720
40	20, 7065	22, 1643	24, 4331	26, 5093	29, 0505	51, 8050	55, 7585	59, 3417	63, 6907	66, 7659
50	27, 9907	29, 7067	32, 3574	34, 7042	37, 6886	63, 1671	67, 5048	71, 4202	76, 1539	79, 4900
60	35, 5346	37, 4848	40, 4817	43, 1879	46, 4589	74, 3970	79, 0819	83, 2976	88, 3794	91, 9517
70	43, 2752	45, 4418	48, 7576	51, 7393	55, 3290	85, 5271	90, 5312	95, 0231	100, 425	104, 215
80	51, 1720	53, 5400	57, 1532	60, 3915	64, 2778	96, 5782	101, 879	106, 629	112, 329	116, 321
90	59, 1953	61, 7541	65, 6466	69, 1260	73, 2912	107, 565	113, 145	118, 136	124, 116	128, 299
100	67, 3276	70, 0648	74, 2219	77, 9295	82, 3581	118, 498	124, 342	129, 561	135, 807	140, 169

Таблица П1.5. Значения $100Q\%$ -ных точек $v_Q^2(\nu_1, \nu_2)$ F -распределения с числом степеней свободы числителя ν_1 и знаменателя ν_2 .

ν_2	ν_1																			$Q=0.1$	$Q=0.05$	$Q=0.01$	$Q=0.001$
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞				
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33				
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49				
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13				
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76				
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10				
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72				
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47				
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29				
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16				
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06				
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97				
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90				
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85				
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80				
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76				
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72				
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69				
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66				
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63				
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61				
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59				
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57				
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55				
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53				

25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

Q=0.05

1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.60	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13

Продолжение таблицы III.5

v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,05	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

	$Q=0,01$																		
1	4,052	4,099,5	5,403	5,625	5,764	5,859	5,928	5,982	6,022	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46

5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,05	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Примечание. При вычислении 100%-ных точек для $Q \geq 0,9$ следует использоватья тождеством $v_Q^2(u_1, u_2) = (v_{1-Q}^2(u_2, u_1))^{-1}$.

Таблица П1.6. Значения 100Q%-ных точек $t_Q(\nu)$ распределения Стьюдента с ν степенями свободы.

ν	$Q=0.4$ $2Q=0.8$	0.25 0.5	0.1 0.2	0.05 0.1	0.025 0.05	0.01 0.02	0.005 0.01	0.0025 0.005
1	0,325	1,000	3,078	6,314	12,706	31,821	63,657	127,32
2	289	0,816	1,886	2,920	4,303	6,965	9,925	14,089
3	277	765	1,638	2,353	3,182	4,541	5,841	7,453
4	271	741	1,533	2,132	2,776	3,747	4,604	5,598
5	0,267	0,727	1,476	2,015	2,571	3,365	4,032	4,773
6	265	718	1,440	1,943	2,447	3,143	3,707	4,317
7	263	711	1,415	1,895	2,365	2,998	3,499	4,029
8	262	706	1,397	1,860	2,306	2,896	3,355	3,833
9	261	703	1,383	1,833	2,262	2,821	3,250	3,690
10	0,260	0,700	1,372	1,812	2,228	2,764	3,169	3,581
11	260	697	1,363	1,796	2,201	2,718	3,106	3,497
12	259	695	1,356	1,782	2,179	2,681	3,055	3,428
13	259	694	1,350	1,771	2,160	2,650	3,012	3,372
14	258	692	1,345	1,761	2,145	2,624	2,977	3,326
15	0,258	0,691	1,341	1,753	2,131	2,602	2,947	3,286
16	258	690	1,337	1,746	2,120	2,583	2,921	3,252
17	257	689	1,333	1,740	2,110	2,567	2,898	3,222
18	257	688	1,330	1,734	2,101	2,552	2,878	3,197
19	257	688	1,328	1,729	2,093	2,539	2,861	3,174
20	0,257	0,687	1,325	1,725	2,086	2,528	2,845	3,153
21	257	686	1,323	1,721	2,080	2,518	2,831	3,135
22	256	686	1,321	1,717	2,074	2,508	2,819	3,119
23	256	685	1,319	1,714	2,069	2,500	2,807	3,104
24	256	685	1,318	1,711	2,064	2,492	2,797	3,091
25	0,256	0,684	1,316	1,708	2,060	2,485	2,787	3,078
26	256	684	1,315	1,706	2,056	2,479	2,779	3,067
27	256	684	1,314	1,703	2,052	2,473	2,771	3,057
28	256	683	1,313	1,701	2,048	2,467	2,763	3,047
29	256	683	1,311	1,699	2,045	2,462	2,756	3,038
30	0,256	0,683	1,310	1,697	2,042	2,457	2,750	3,030
40	255	681	1,303	1,684	2,021	2,423	2,704	2,971
60	254	679	1,296	1,671	2,000	2,390	2,660	2,915
120	254	677	1,289	1,658	1,980	2,358	2,617	2,860
∞	253	674	1,282	1,645	1,960	2,326	2,576	2,807

Таблица П1.7. Преобразование Фишера (z -преобразование) выборочного коэффициента корреляции $\hat{r}$ ($z = \operatorname{arcthr}\hat{r}$).

$\hat{r}$	,000	,002	,004	,006	,008	$\hat{r}$	,000	,002	,004	,006	,008
0,00	0000	0020	0040	0060	0080	0,50	5493	5520	5547	5573	5600
1	0100	0120	0140	0160	0180	1	5627	5654	5682	5709	5736
2	0200	0220	0240	0260	0280	2	5763	5791	5818	5846	5874
3	0300	0320	0340	0360	0380	3	5901	5929	5957	5985	6013
4	0400	0420	0440	0460	0480	4	6042	6070	6098	6127	6155
0,05	0500	0520	0541	0561	0581	0,55	6184	6213	6241	6270	6299
6	0601	0621	0641	0661	0681	6	6328	6358	6387	6416	6446
7	0701	0721	0741	0761	0782	7	6475	6505	6535	6565	6595
8	0802	0822	0842	0862	0882	8	6625	6655	6685	6716	6746
9	0902	0923	0943	0963	0983	9	6777	6807	6838	6869	6900
0,10	1003	1024	1044	1064	1084	0,60	6931	6963	6994	7026	7057
1	1104	1125	1145	1165	1186	1	7089	7121	7153	7185	7218
2	1206	1226	1246	1267	1287	2	7250	7283	7315	7348	7381
3	1307	1328	1348	1368	1389	3	7414	7447	7481	7514	7548
4	1409	1430	1450	1471	1491	4	7582	7616	7650	7684	7718
0,15	1511	1532	1552	1573	1593	0,65	7753	7788	7823	7858	7893
6	1614	1634	1655	1676	1696	6	7928	7964	7999	8035	8071
7	1717	1737	1758	1779	1799	7	8107	8144	8180	8217	8254
8	1820	1841	1861	1882	1903	8	8291	8328	8366	8404	8441
9	1923	1944	1965	1986	2007	9	8480	8518	8556	8595	8634
0,20	2027	2048	2069	2090	2111	0,70	8673	8712	8752	8792	8832
1	2132	2153	2174	2195	2216	1	8872	8912	8953	8994	9035
2	2237	2258	2279	2300	2321	2	9076	9118	9160	9202	9245
3	2342	2363	2384	2405	2427	3	9287	9330	9373	9417	9461
4	2448	2469	2490	2512	2533	4	9505	9549	9594	9639	9684
0,25	2554	2575	2597	2618	2640	0,75	0,973	0,978	0,982	0,987	0,991
6	2661	2683	2704	2726	2747	6	0,996	1,001	1,006	1,011	1,015
7	2769	2790	2812	2833	2855	7	1,020	1,025	1,030	1,035	1,040
8	2877	2899	2920	2942	2964	8	1,045	1,050	1,056	1,061	1,066
9	2986	3008	3029	3051	3073	9	1,071	1,077	1,082	1,088	1,093
0,30	3095	3117	3139	3161	3183	0,80	1,099	1,104	1,110	1,116	1,121
1	3205	3228	3250	3272	3294	1	1,127	1,133	1,139	1,145	1,151
2	3316	3339	3361	3383	3406	2	1,157	1,163	1,169	1,175	1,182
3	3428	3451	3473	3496	3518	3	1,188	1,195	1,201	1,208	1,214
4	3541	3564	3586	3609	3632	4	1,221	1,228	1,235	1,242	1,249
0,35	3654	3677	3700	3723	3746	0,85	1,256	1,263	1,271	1,278	1,286
6	3769	3792	3815	3838	3861	6	1,293	1,301	1,309	1,317	1,325
7	3884	3907	3931	3954	3977	7	1,333	1,341	1,350	1,358	1,367
8	4001	4024	4047	4071	4094	8	1,376	1,385	1,394	1,403	1,412
9	4118	4142	4165	4189	4213	9	1,422	1,432	1,442	1,452	1,462

Продолжение табл. П1.7.

$\hat{r}$	,000	,002	,004	,006	,008	$\hat{r}$	,000	,002	,004	,006	,008
0,40	4236	4260	4284	4308	4332	0,90	1,472	1,483	1,494	1,505	1,516
1	4356	4380	4404	4428	4453	1	1,528	1,539	1,551	1,564	1,576
2	4477	4501	4526	4550	4574	2	1,589	1,602	1,616	1,630	1,644
3	4599	4624	4648	4673	4698	3	1,658	1,673	1,689	1,705	1,721
4	4722	4747	4772	4797	4822	4	1,738	1,756	1,774	1,792	1,812
0,45	4847	4872	4897	4922	4948	0,95	1,832	1,853	1,874	1,897	1,921
6	4973	4999	5024	5049	5075	6	1,946	1,972	2,000	2,029	2,060
7	5101	5126	5152	5178	5204	7	2,092	2,127	2,165	2,205	2,249
8	5230	5256	5282	5308	5334	8	2,298	2,351	2,410	2,477	2,555
9	5361	5387	5413	5440	5466	9	2,647	2,759	2,903	3,106	3,453
$\hat{r}$	,000	,002	,004	,006	,008	$\hat{r}$	,000	,002	,004	,006	,008

Указание. При работе с отрицательными значениями $\hat{r}$ и $\hat{z}$ используйте свойство нечетности функций $\text{th } \hat{z}$ и $\text{arcth } \hat{r}$, т.е. $\text{arcth } (-\hat{r}) = -\text{arcth } \hat{r}$.

П р и м е р ы.

1. Дано $\hat{r} = 0,206$. Определить $z = \text{arcth } 0,206$.

Находим (в левом столбце таблицы) строку, соответствующую $\hat{r} = 0,20$. Чтобы получить заданное значение $\hat{r}$, к 0,20 надо прибавить 0,006, а потому искомое число находится (в этой строке) в столбце, расположенном под 0,006. Итак, $z = \text{arcth } 0,2090$.

2. Дано $\hat{r} = -0,515$. Определить $z = \text{arcth } (-0,515)$.

Находим (в левом столбце таблицы) строку, соответствующую $\hat{r} = 0,51$. Чтобы получить значение $\hat{r} = 0,515$, к 0,51 надо прибавить 0,005, а потому $\text{arcth } 0,515$ находится как среднее арифметическое двух чисел данной строки, расположенных в столбцах, соответствующих верхним индексам 0,006 и 0,004, т.е.

$$\text{arcth } 0,515 = \frac{0,5709 + 0,5682}{2} = 0,56955.$$

Соответственно, $z = \text{arcth } (-0,515) = -\text{arcth } 0,515 = -0,56955$.

3. Дано $z = 0,8752$. Определить $\hat{r} = \text{th } z$.

Находим в таблице число, равное 0,8752, и определяем, какому значению $\hat{r}$ оно соответствует. В нашем случае $\hat{r} = 0,704$.

П р и м е ч а н и е. В тех случаях, когда в таблице не найдется в точности заданного, берут два приближенных (ближайших к нему) значения — с недостатком и с избытком. Искомое значение $\hat{r}$ будет лежать между двумя значениями $\hat{r}_1$ и $\hat{r}_2$, соответствующими этим приближенным величинам z .

Таблица П1.8. Верхняя граница доверительного интервала для истинного значения коэффициента корреляции при условии отсутствия линейной корреляционной связи (при доверительной вероятности $P = 1 - 2Q$).

n — 2	Q					
	0,05	0,025	0,01	0,005	0,0025	0,0005
1	0,9877	0,92692	0,9507	0,93877	0,94692	0,94877
2	9000	9500	9800	92000	9500	92000
3	805	878	9343	9587	9740	92114
4	729	811	882	9172	9417	9741
5	669	754	833	875	9056	9509
6	0,621	0,707	0,789	0,834	0,870	0,9249
7	582	666	750	798	836	898
8	549	632	715	765	805	872
9	521	602	685	735	776	847
10	497	576	658	708	750	823
11	0,476	0,553	0,634	0,684	0,726	0,801
12	457	532	612	661	703	780
13	441	514	592	641	683	760
14	426	497	574	623	664	742
15	412	482	558	606	647	725
16	0,400	0,468	0,543	0,590	0,631	0,708
17	389	456	529	575	616	693
18	378	444	516	561	602	679
19	369	433	503	549	589	665
20	360	423	492	537	576	652
25	0,323	0,381	0,445	0,487	0,524	0,597
30	296	349	409	449	484	554
35	275	325	381	418	452	519
40	257	304	358	393	425	490
45	243	288	338	372	403	465
50	0,231	0,273	0,322	0,354	0,384	0,443
60	211	250	295	325	352	408
70	195	232	274	302	327	380
80	183	217	257	283	307	357
90	173	205	242	267	290	338
100	164	195	230	254	276	321

П р и м е ч а н и е. Верхний индекс (2,3 и т. д.) над цифрой 9 означает, что эта цифра занимает первые 2,3 и т. д. разряда десятичной дроби. Например, $0,9^{2692} = 0,9999692$.

П р и м е р. Если мы оцениваем корреляционную связь по $n = 20$ наблюдениям, то при доверительной вероятности $P = 0,95$ (т. е. при $2Q = 0,05$) значение коэффициента корреляции, не превосходящее по абсолютной величине 0,444, еще не говорит о статистической значимости этой корреляционной связи (т. е. о том, что истинное значение коэффициента корреляции ρ отлично от нуля).

Таблица П1.9.

Проверка статистической значимости корреляционной связи с помощью рангового коэффициента корреляции Спирмена $\tau(S)$

n = 4		n = 5		n = 6		n = 7		n = 8		n = 9		n = 10	
S_C	q	S_C	q	S_C	q	S_C	q	S_C	q	S_C	q	S_C	q
12	0,458	22	0,475	50	0,210	74	0,249	108	0,250	156	0,218	208	0,235
14	375	24	392	52	178	78	198	114	195	164	168	218	184
16	208	26	342	54	149	82	151	120	150	172	125	228	139
18	167	28	258	56	121	86	118	126	108	180	089	238	102
20	042	30	225	58	088	90	083	132	076	188	060	248	072
		32	0,175	60	0,068	94	0,055	138	0,048	196	0,038	258	0,048
		34	117	62	051	98	033	144	029	204	022	268	030
		36	067	64	029	102	017	150	014	212	011	278	017
		38	042	66	017	106	0062	156	0054	220	0041	288	0087
		40	0083	68	0083	110	0014	162	0011	228	0010	298	0036
				70	0,0014							308	0,001
20		40		70		112		168		240		330	

Таблица П1.10.

Проверка статистической значимости корреляционной связи с помощью рангового коэффициента корреляции Кендалла $\tau^{(K)}$

S_K	n				S_K	n		
	4	5	8	9		6	7	10
0	0,625	0,592	0,548	0,540	1	0,500	0,500	0,500
2	375	408	452	460	3	360	386	431
4	167	242	360	381	5	235	281	364
6	042	117	274	306	7	136	191	300
8		042	199	238	9	068	119	242
10		0,0083	0,138	0,179	11	0,028	0,068	0,190
12			089	130	13	0083	035	146
14			054	090	15	0014	015	108
16			031	060	17		0054	078
18			016	038	19		0014	054
20			0,0071	0,022	21		0,0002	0,036
22			0028	012	23			023
24			0009	0063	25			014
26			0002	0029	27			0083
28				0012	29			0046
30				0,0004	31			0,0023
					33			0011

Продолжение

Вероятность того, что данное значение S будет достигнуто или превзойдено, для $n=4$, $m=3$ и $m=5$

S	$m=3$	$m=5$	S	$m=5$
1	1,000	1,000	61	0,055
3	0,958	0,975	65	0,044
5	0,910	0,944	67	0,034
9	0,727	0,857	69	0,031
11	0,608	0,771	73	0,023
13	0,524	0,709	75	0,020
17	0,446	0,652	77	0,017
19	0,342	0,561	81	0,012
21	0,300	0,521	83	0,0087
25	0,207	0,445	85	0,0067
27	0,175	0,408	89	0,0055
29	0,148	0,372	91	0,0031
33	0,075	0,298	93	0,0023
35	0,054	0,260	97	0,0018
37	0,033	0,226	99	0,0016
41	0,017	0,210	101	0,0014
43	0,0017	0,162	105	0,00064
45	0,0017	0,141	107	0,00033
49		0,123	109	0,00021
51		0,107	113	0,00014
53		0,093	117	0,000048
57		0,075	125	0,000030
59		0,067		

Продолжение

Вероятность того, что данное значение S будет достигнуто или превзойдено, для $n=4$ и $m=2$, $m=4$ и $m=6$

S	$m=2$	$m=4$	$m=6$	S	$m=6$
0	1,000	1,000	1,000	82	0,035
2	0,958	0,992	0,996	84	0,032
4	0,833	0,928	0,957	86	0,029
6	0,792	0,900	0,940	88	0,023
8	0,625	0,800	0,874	90	0,022
10	0,542	0,754	0,844	94	0,017
12	0,458	0,677	0,789	96	0,014
14	0,375	0,649	0,772	98	0,013
16	0,208	0,524	0,679	100	0,010
18	0,167	0,508	0,668	102	0,0096
20	0,042	0,432	0,609	104	0,0085
22		0,389	0,574	106	0,0073
24		0,355	0,541	108	0,0061

Продолжение Таблицы П1.11^а

S	$m=2$	$m=4$	$m=6$	S	$m=6$
26		0,324	0,512	110	0,0057
30		0,242	0,431	114	0,0040
32		0,200	0,386	116	0,0033
34		0,190	0,375	118	0,0028
36		0,158	0,338	120	0,0023
38		0,141	0,317	122	0,0020
40		0,105	0,270	126	0,0015
42		0,094	0,258	128	0,0090
44		0,077	0,230	130	0,0087
46		0,068	0,218	132	0,0073
48		0,054	0,197	134	0,0065
50		0,052	0,194	136	0,0040
52		0,036	0,163	138	0,0036
54		0,033	0,155	140	0,0028
56		0,019	0,127	144	0,0024
58		0,014	0,114	146	0,0022
62		0,012	0,108	148	0,0012
64		0,0069	0,089	150	0,0095
66		0,0062	0,088	152	0,0062
68		0,0027	0,073	154	0,0046
70		0,0027	0,066	158	0,0024
72		0,0016	0,060	160	0,0016
74		0,0094	0,056	162	0,0012
76		0,0094	0,043	164	0,0080
78		0,0094	0,041	170	0,0024
80		0,0072	0,037	180	0,0013

Продолжение

Вероятность того, что данное значение S будет достигнуто или превзойдено, для $n=5$ и $m=3$

S	$m=3$	S	$m=3$	S	$m=3$	S	$m=3$
0	1,000	22	0,649	44	0,236	66	0,038
2	1,000	24	0,595	46	0,213	68	0,028
4	0,988	26	0,559	48	0,172	70	0,026
6	0,972	28	0,493	50	0,163	72	0,017
8	0,941	30	0,475	52	0,127	74	0,015
10	0,914	32	0,432	54	0,117	76	0,0078
12	0,845	34	0,406	56	0,096	78	0,0053
14	0,831	36	0,347	58	0,080	80	0,0040
16	0,768	38	0,326	60	0,063	82	0,0028
18	0,720	40	0,291	62	0,056	86	0,0090
20	0,682	42	0,253	64	0,045	90	0,0069

Таблица П1.11⁶. Критические значения статистики S при уровне значимости $\alpha = 0,05$ для проверки статистической значимости выборочного значения коэффициента конкордации $\widehat{W}(m)$.

m	n					Дополнительные значения для $n = 3$	
	3	4	5	6	7	m	S
3			64,4	103,9	157,3	9	54,0
4		49,5	88,4	143,3	217,0	12	71,9
5		62,6	112,3	182,4	276,2	14	83,8
6		75,7	136,1	221,4	335,2	16	95,8
8	48,1	101,7	183,7	299,0	453,1	18	107,7
10	60,0	127,8	231,2	376,7	571,0		
15	89,8	192,9	349,8	570,5	864,9		
20	119,7	258,0	468,5	764,4	1158,7		

ПРИЛОЖЕНИЕ 2. НЕОБХОДИМЫЕ СВЕДЕНИЯ ИЗ МАТРИЧНОЙ АЛГЕБРЫ

Все модели и методы эконометрики, эксплуатирующие понятия и приемы *многомерного статистического анализа*, — а это и классификация многомерных данных, и всевозможные конструкции факторного анализа, включая метод главных компонент, и множественная регрессия, и, конечно, системы одновременных уравнений, и все, что связано с решением систем алгебраических уравнений (а сюда относятся и корреляционный анализ, и анализ временных рядов), — *требуют, для своего компактного и эффективного описания и анализа, использования аппарата матричной алгебры*. И если при **описании** этих моделей мы еще, как правило, можем обойтись без такого аппарата (правда, за счет утомительных и чрезвычайно громоздких выкладок и выражений, приводящих к *покомпонентной* записи требуемых соотношений), то **при анализе их свойств и выводе связанных с этим математических результатов** без использования методов линейной алгебры мы вынуждены были бы в ряде ситуаций просто отступить, не добившись желаемого.

Поэтому специалисту (или студенту, готовящемуся стать таковым), предполагающему всерьез работать с эконометрическими методами и моделями, необходимо внутренне настроиться на неизбежность расходования некоторого времени и интеллектуальных усилий на овладение определенным минимумом сведений из матричной алгебры.

При описании этих сведений мы будем следовать обозначениям, принятым в данном учебнике, а именно: буквы, набранные *жирным* шрифтом, будут обозначать *матрицы*, а буквы *прописные* будут использоваться для обозначения *векторов* (представляющих собой, правда, важный частный случай тех же матриц). Доказательства приводимых здесь фактов из линейной алгебры будут, как правило, опускаться.

П2.1. Виды матриц и действия с ними

Матрицей называется *прямоугольная таблица чисел (элементов) вида*

$$A = (a_{ij})_{\substack{i=1,2,\dots,n; \\ j=1,2,\dots,m}} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix}. \quad (\text{П2.1})$$

В общем обозначении элемента a_{ij} матрицы A первый нижний индекс (i) указывает номер той строки, а второй нижний индекс (j) — номер того столбца, на пересечении которых находится этот элемент. Если матрица A содержит nm элементов, расположенных на n строках и m столбцах, то говорят, что мы имеем $n \times m$ -матрицу A или матрицу A размерности $n \times m$.

Квадратной матрицей называется матрица, у которой число строк равно числу столбцов, т.е. $n \times n$ — матрица (при любом целом $n \geq 1$). Элементы $a_{11}, a_{22}, \dots, a_{nn}$ квадратной матрицы образуют ее *главную диагональ*.

Среди квадратных матриц выделим:

- **диагональную матрицу** D , у которой все элементы, кроме элементов, стоящих на главной диагонали, равны нулю, т.е.

$$D = \begin{pmatrix} d_{11} & 0 & 0 & \dots & 0 & 0 \\ 0 & d_{22} & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & d_{nn} \end{pmatrix} = \text{diag}(d_{11}, d_{22}, \dots, d_{nn});$$

заметим, что ковариационная матрица $\Sigma = (\sigma_{ij})$ любой последовательности $\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(n)}$ взаимно не коррелированных случайных величин является диагональной, т.к. $\sigma_{ij} = E[(\xi^{(i)} - E\xi^{(i)})(\xi^{(j)} - E\xi^{(j)})] = 0$ при любых $i \neq j$.

- **единичную матрицу** I_n , которая является частным случаем $n \times n$ диагональной матрицы, поскольку все ее диагональные элементы равны единице:

$$I_n = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix};$$

нижний индекс n определяет размерность матрицы, и в тех случаях, когда эта размерность очевидна из контекста, он может опускаться.

- **симметрическую (симметричную)** матрицу, а именно такую матрицу, у которой элементы, симметрично расположенные относительно главной диагонали, равны между собой, т.е. $a_{ij} = a_{ji}$ для всех $i, j = 1, 2, \dots, n$; заметим, что всякая ковариационная матрица $\Sigma = (\sigma_{ij})$ является симметрической по определению, т.к. $\sigma_{ij} = E[(\xi^{(i)} - E\xi^{(i)})(\xi^{(j)} - E\xi^{(j)})] = E[(\xi^{(j)} - E\xi^{(j)})(\xi^{(i)} - E\xi^{(i)})] = \sigma_{ji}$.

Равенство матриц. Две матрицы $A = (a_{ij})$ и $B = (b_{ij})$ называются равными, если они имеют одну и ту же размерность и если $a_{ij} = b_{ij}$ для

всех i и j . Это означает, что равные матрицы *совпадают поэлементно*.

Вектор-строка — это матрица размерности $1 \times m$ ($m \geq 1$), т.е. *матрица, состоящая из единственной строки длины m* . Например, результаты наблюдения значений p анализируемых переменных $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ на одном объекте, зарегистрированные в определенный момент времени t , образуют вектор-столбец

$$X_t = (x_t^{(1)}, x_t^{(2)}, \dots, x_t^{(p)}). \quad (\text{П2.2})$$

Вектор-столбец — это матрица размерности $n \times 1$ ($n \geq 1$), т.е. *матрица, состоящая из единственного столбца длины n* . Например, результаты наблюдения одной какой-либо переменной $x^{(j)}$ на n статистически обследованных объектах (или на одном объекте, но зарегистрированные в n последовательных моментов времени) можно представить в виде вектора-столбца

$$X^{(j)} = \begin{pmatrix} x_1^{(j)} \\ x_2^{(j)} \\ \vdots \\ x_n^{(j)} \end{pmatrix}. \quad (\text{П2.3})$$

Чтобы выделить эти специальные частные случаи матриц, мы будем обозначать векторы-строки и векторы-столбцы *прописными* буквами алфавита, а их компоненты — *теми же самыми, но строчными буквами*.

Условимся обозначать в дальнейшем вектор-строку длины m , состоящую из одних нулей, с помощью $0_{1,m}$ (или просто 0_m , если из контекста ясно, что речь идет о строках). Аналогично вектор-столбец из нулей длины n будем обозначать $0_{n,1}$ (или просто 0_n).

Перейдем к описанию *действий над матрицами*.

Транспонирование матрицы A определяется как действие, в результате которого из A получается новая матрица A^T , строками которой служат столбцы матрицы A , а столбцами — строки матрицы A при сохранении их порядка. Таким образом, первая строка матрицы A становится первым столбцом матрицы A^T , вторая строка матрицы A — вторым столбцом матрицы A^T и т.д., так что (i, j) -й элемент a_{ij} матрицы A становится (j, i) -м элементом матрицы A^T . Так, при построении регрессионных моделей мы оперировали как с $n \times (p+1)$ -матрицей наблюдений

объясняющих переменных

$$X = \begin{pmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(p)} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(p)} \end{pmatrix}, \quad (\text{П2.4})$$

так и с *транспонированной* $(p+1) \times n$ -матрицей

$$X^T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ \dots & \dots & \dots & \dots \\ x_1^{(p)} & x_2^{(p)} & \dots & x_n^{(p)} \end{pmatrix}. \quad (\text{П2.5})$$

Заметим, что из определений операции транспонирования и свойства симметричности матрицы следует непосредственно, что *транспонирование не меняет симметрическую матрицу*, т. е. если A — симметрична, то $A^T = A$ и обратно: если $A^T = A$, то матрица A квадратна и симметрична.

Очевидно также, что последовательное двукратное применение к матрице A операции транспонирования приводит к *исходной* матрице A , т. е. $(A^T)^T = A$.

Сложение двух матриц. Если A и B -матрицы одной размерности, то можно определить новую матрицу $C = A + B$, которая будет иметь ту же размерность, что и A и B , а ее элементы c_{ij} определяются для всех i и j соотношениями $c_{ij} = a_{ij} + b_{ij}$. Заметим, что из определений операций транспонирования и сложения непосредственно следует, что

$$(A + B)^T = A^T + B^T. \quad (\text{П2.6})$$

Умножение матрицы на число. Произведение матрицы A на число λ определяется как

$$\lambda A = (\lambda a_{ij})_{\substack{i=1,2,\dots,n \\ j=1,2,\dots,m}},$$

т. е. каждый элемент матрицы A умножается на это число.

Произведение матриц. Если число столбцов $n \times m$ -матрицы A равно числу строк $m \times k$ -матрицы B , то может быть определена операция умножения AB матрицы A на матрицу B (при этом говорят, что «матрица B умножается на матрицу A слева»). Элементы c_{ij} такого произведения $C = AB$ определяются для всех $i = 1, 2, \dots, n$ и $j = 1, 2, \dots, k$

соотношениями

$$c_{ij} = \sum_{l=1}^m a_{il}b_{lj}. \quad (\text{П2.7})$$

Таким образом, (i, j) -й элемент произведения C вычисляется как *скалярное произведение* i -й строки матрицы A и j -го столбца матрицы B , т. е. как сумма попарных произведений элементов i -й строки первой матрицы на соответствующие (т. е. стоящие на тех же по порядку местах) элементы j -го столбца второй матрицы. При этом автоматически определяется размерность матрицы-произведения C : она будет иметь столько же строк (n), сколько имел первый сомножитель, и столько же столбцов (k), сколько их имел второй сомножитель.

Заметим, что при перестановке сомножителей произведение BA может просто не существовать, но даже если оно существует (как это и бывает в случае, если матрицы A и B — квадратные и одной размерности или если сомножители имеют размерности $n \times m$ и $m \times n$), то, *вообще говоря, коммутативный закон для умножения матриц не имеет места, т. е.*

$$AB \neq BA. \quad (\text{П2.8})$$

Отметим еще несколько полезных для эконометрических приложений свойств произведения матриц.

Ассоциативный закон:

$$ABC = (AB)C = A(BC). \quad (\text{П2.9})$$

Дистрибутивный закон:

$$A(B + C) = AB + AC, \quad (B + C)A = BA + CA. \quad (\text{П2.10})$$

Умножение на единичную матрицу: для любой $n \times n$ -матрицы A имеют место тождества

$$AI_n = I_n A = A. \quad (\text{П2.11})$$

Транспонирование произведения матриц:

$$(A_1 A_2 \dots A_k)^T = A_k^T A_{k-1}^T \dots A_2^T A_1^T \quad (\text{П2.12})$$

(доказывается индукцией: непосредственным сравнением элементов матриц $(A_1 A_2)^T$ и $A_2^T A_1^T$ доказывается справедливость (П2.12) для случая $k = 2$ и т. д.).

Произведения вида AA^T и $A^T A$ играют заметную роль в эконометрических построениях. Так, произведение $X^T X$, в котором матрица X определена соотношением (П2.4), является неизменным «участником» всех

основных формул классического метода наименьших квадратов. Заметим, что если A — матрица размерности $n \times m$, то произведение AA^T будет иметь размерность $n \times n$, в то время как произведение $A^T A$ — это матрица размерности $m \times m$. Но в любом случае *произведения вида AA^T и $A^T A$ всегда являются квадратными симметрическими матрицами.*

Обратим внимание на специальный случай, когда $X = (x_1, x_2, \dots, x_n)^T$ — это вектор-столбец, состоящий из n элементов (или *вектор-столбец длины n*). Тогда

$$X^T X = \sum_{i=1}^n x_i^2 \quad (П2.13)$$

это 1×1 -матрица, т. е. *число*, а

$$XX^T = \begin{pmatrix} x_1^2 & x_1 x_2 & \dots & x_1 x_n \\ x_2 x_1 & x_2^2 & \dots & x_2 x_n \\ \dots & \dots & \dots & \dots \\ x_n x_1 & x_n x_2 & \dots & x_n^2 \end{pmatrix} \quad (П2.14)$$

это матрица размерности $n \times n$.

Матрицы типа (П2.13) и (П2.14) играют заметную роль в регрессионном анализе и в системах одновременных уравнений (см. гл. 14, 15, 17). Действительно, если в качестве компонент x_i вектора X рассмотреть регрессионные «невязки» $y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i^{(1)} - \dots - \hat{\theta}_p x_i^{(p)}$ ($i = 1, 2, \dots, n$), то произведение (П2.13) даст нам сумму квадратов «невязок», которая играет важную роль в анализе точности регрессионной модели. Если же в качестве компонент x_i вектора X рассмотреть отклонения i -й объясняющей переменной $x^{(i)}$ от своего среднего значения $a^{(i)}$ ($i = 1, 2, \dots, p$), то произведение (П2.14) после применения к нему операции усреднения (математического ожидания) E даст $p \times p$ -ковариационную матрицу объясняющих переменных Σ_X .

Прямое (или кронекерово) произведение $n \times m$ -матрицы A и $k \times l$ -матрицы B обозначается $A \otimes B$, имеет размерность $nk \times ml$ и подсчитывается по формуле

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1m}B \\ a_{21}B & a_{22}B & \dots & a_{2m}B \\ \dots & \dots & \dots & \dots \\ a_{n1}B & a_{n2}B & \dots & a_{nm}B \end{pmatrix}, \quad (П2.15)$$

где правая часть представляет собой матрицу, *составленную из блоков вида $a_{ij}B$* . Каждый такой блок сам является матрицей и, как легко видеть,

имеет размерность $k \times l$. Повторенные n раз по строкам, они дают общее количество строк, равное nk , а повторенные m раз по столбцам, они дают общее количество столбцов, равное ml (подробнее о *матрицах блочного типа*, или о *составленных матрицах*, так же, как и свойствах прямого произведения, см. ниже в п. П2.8).

Отметим, что некоторые эконометрические построения (в первую очередь, связанные с вычислительными проблемами *систем регрессионных уравнений*, см. гл. 4 в томе 2) значительно упрощаются, если использовать понятие прямого (кронекерова) произведения.

Матричное дифференцирование. Достаточно полное изложение аппарата векторного и матричного дифференциального исчисления читатель найдет в книге [Magnus J.R., Neudecker H.]¹. Мы же остановимся лишь на тех определениях и результатах, которые непосредственно использованы в нашем учебнике.

Пусть $\Theta = (\theta_1, \theta_2, \dots, \theta_m)^\top$ — $m \times 1$ -матрица (т.е. вектор-столбец длины m), компоненты которой играют роль неизвестных параметров эконометрической модели, а $A(\Theta) = (a_1(\Theta), a_2(\Theta), \dots, a_n(\Theta))^\top$ — $n \times 1$ -матрица (т.е. вектор-столбец длины n), компоненты которой интерпретируются как некоторые характеристики этой модели, зависящие от Θ .

Производной $n \times 1$ -векторной функции $A(\Theta)$ по $m \times 1$ -векторному аргументу Θ называется $m \times n$ -матрица

$$\frac{\partial A(\Theta)}{\partial \Theta} = \begin{pmatrix} \frac{\partial a_1(\Theta)}{\partial \theta_1} & \frac{\partial a_1(\Theta)}{\partial \theta_2} & \dots & \frac{\partial a_1(\Theta)}{\partial \theta_m} \\ \frac{\partial a_2(\Theta)}{\partial \theta_1} & \frac{\partial a_2(\Theta)}{\partial \theta_2} & \dots & \frac{\partial a_2(\Theta)}{\partial \theta_m} \\ \dots & \dots & \dots & \dots \\ \frac{\partial a_n(\Theta)}{\partial \theta_1} & \frac{\partial a_n(\Theta)}{\partial \theta_2} & \dots & \frac{\partial a_n(\Theta)}{\partial \theta_m} \end{pmatrix}. \quad (\text{П2.16})$$

Матрицы вида (П2.16) используются, в частности, в качестве матриц преобразований («якобианов») в теории преобразований случайных величин (см., например, п. 4.4).

Рассмотрим важные для эконометрических приложений частные случаи функции $A(\Theta)$:

1⁰. $A(\Theta) = X\Theta$, где X — симметрическая матрица размерности $n \times m$.
Тогда $\frac{\partial A(\Theta)}{\partial \Theta} = \frac{\partial (X\Theta)}{\partial \Theta} = X$.

2⁰. $A(\Theta) = \Theta^\top B \Theta$, где B — симметрическая квадратная матрица размерности $m \times m$.

Тогда $\frac{\partial A(\Theta)}{\partial \Theta} = \frac{\partial (\Theta^\top B \Theta)}{\partial \Theta} = 2\Theta^\top B$.

¹ Готовится русский перевод этой книги.

3°. $A(\Theta) = X^T \Theta$, где X — вектор-столбец длины m .

$$\text{Тогда } \frac{\partial A(\Theta)}{\partial \Theta} = \frac{\partial (X^T \Theta)}{\partial \Theta} = X^T.$$

П2.2. Основные числовые характеристики квадратной матрицы: определитель, след

Любой квадратной матрице A можно сопоставить некоторый набор ее числовых характеристик. В данном пункте остановимся на двух из них. Одна из таких характеристик называется *определителем* (*детерминантом*) матрицы A и обозначается $\det A$ или $|A|$, а другая определяется как *след матрицы* A и обозначается $\text{tr}(A)$.

Определитель (детерминант) $n \times n$ -матрицы A вычисляется по формуле

$$\det A = \sum_{j_1=1}^n \sum_{j_2=1}^n \dots \sum_{j_n=1}^n (-1)^{\nu(j_1, j_2, \dots, j_n)} a_{1j_1} a_{2j_2} \dots a_{nj_n}, \quad (\text{П2.17})$$

где суммирование ведется по всем возможным комбинациям *различных* столбцов (т.е. по всем возможным перестановкам вторых индексов), а $\nu(j_1, j_2, \dots, j_n)$ — это минимальное число инверсий (т.е. парных обменов местами), которое надо совершить с элементами *исходной* перестановки $(1, 2, \dots, n)$, чтобы получить перестановку $(j_1, j_2, \dots, j_n)$. Очевидно, общее число слагаемых в правой части (П2.17) составит при таком определении n -факториал $(n!)$.

Для малых размерностей матриц это определение приводит, в частности, к следующим результатам:

$$\text{а) } n = 2: \quad \det A = \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21};$$

$$\begin{aligned} \text{б) } n = 3: \quad \det A &= \det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \\ &= a_{11}a_{22}a_{33} + a_{13}a_{21}a_{32} + a_{12}a_{23}a_{31} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}. \end{aligned}$$

Приведем здесь важнейшие свойства определителя:

- 1) $\det(A B) = \det(B A) = \det A \cdot \det B$;
- 2) $\det(\lambda A) = \lambda^n \det A$ (λ — число, n — размерность матрицы A);
- 3) $\det[\text{diag}(a_{11}, a_{22}, \dots, a_{nn})] = a_{11}a_{22} \dots a_{nn}$;
- 4) $\det I_n = 1$;
- 5) $\det A^T = \det A$;

- 6) $\det A = 0$, если в матрице A есть две одинаковые строки (два одинаковых столбца);
- 7) инверсия (обмен местами) двух строк (столбцов) матрицы A приводит к изменению знака ее определителя;
- 8) значение определителя матрицы A не изменится, если к любой его строке (столбцу) добавить линейную комбинацию других строк (столбцов).

Последнее свойство нуждается в пояснении. Обозначим $A_{i_1} = (a_{i_1 1}, a_{i_1 2}, \dots, a_{i_1 m})$ — i_1 -ую строку $n \times m$ матрицы A . Линейной комбинацией строк с номерами $i_1, i_2, \dots, i_k$ называется строка, определяемая в соответствии с правилами действий с матрицами по формуле

$$\lambda_1 A_{i_1} + \lambda_2 A_{i_2} + \dots + \lambda_k A_{i_k},$$

где $\lambda_1, \lambda_2, \dots, \lambda_k$ — некоторые числа. Точно так же определяется линейная комбинация столбцов $A_{.j} = (a_{1j}, a_{2j}, \dots, a_{nj})^T$, $j = j_1, j_2, \dots, j_l$;

- 9) разложение определителя по элементам строки (или столбца).

$$\det A = \sum_{j=1}^n a_{ij} (-1)^{i+j} \det A_{ij},$$

где A_{ij} — $(n-1) \times (n-1)$ -матрица, получающаяся из матрицы A вычеркиванием из нее i -й строки и j -го столбца. Величина $\det A_{ij}$ называется *минором* матрицы A , а величина $(-1)^{i+j} \det A_{ij}$ — *алгебраическим дополнением элемента* a_{ij} в матрице A . Кстати, понятие алгебраического дополнения используется в учебнике, например, при вычислении коэффициента детерминации R^2 и частных коэффициентов корреляции по корреляционной матрице $\mathbf{R}$ исследуемого многомерного признака (см. п. 11.2).

Отметим, что если квадратная матрица имеет отличный от нуля определитель, то она называется невырожденной.

След квадратной $n \times n$ -матрицы A (обозначается $\text{tr } A$, от английского слова «trace») определяется как сумма ее диагональных элементов, т. е.

$$\text{tr } A = a_{11} + a_{22} + \dots + a_{nn}. \quad (\text{П2.18})$$

Основные свойства следа матрицы:

- (i) $\text{tr}(AB) = \text{tr}(BA)$;
- (ii) $\text{tr } I_n = n$;
- (iii) $\text{tr}(\lambda A) = \lambda \text{tr } A$, где λ — число;
- (iv) $\text{tr } A^T = \text{tr } A$;

- (v) $\text{tr}(A + B) = \text{tr}A + \text{tr}B$;
- (vi) в качестве частного случая свойства (i) особо отметим ситуации, в которых роль A играет вектор-столбец $X = (x_1, x_2, \dots, x_n)^T$, а роль B — X^T ; и хотя ни для X , ни для X^T след не определен, их произведения XX^T и X^TX являются *квадратными*, соответственно, $n \times n$ - и 1×1 -матрицами и для них действует правило (i), т. е.

$$\text{tr}(XX^T) = \text{tr}(X^TX),$$

где, обращаем внимание читателя, матрица $X^TX = x_1^2 + x_2^2 + \dots + x_n^2$ — это *число*.

Заметим, что и определитель, и след матрицы имеют четкую вероятностную интерпретацию, например, когда в качестве A рассматривается *ковариационная матрица* Σ_ξ многомерной случайной величины $\xi = (\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(p)})^T$. В этом случае и $\text{tr}\Sigma_\xi$ и $\det \Sigma_\xi$ характеризуют *степень многомерного рассеяния* значений этой случайной величины, а $\det \Sigma_\xi$ называется *обобщенной дисперсией* ξ (см. п. 2.6.6).

П2.3. Обратная матрица и ее свойства

В операциях с *числами* для любого *отличного от нуля* числа a существует число $a^{-1} = 1/a$, которое мы называем *обратным* и которое обладает тем характеристическим свойством, что $aa^{-1} = a^{-1}a = 1$. В матричной алгебре роль единицы, как мы видели, выполняет *единичная матрица* I_n , поскольку при умножении на I_n любой квадратной матрицы размерности $n \times n$ справа и слева эта матрица не меняется (см П2.11).

Поэтому по аналогии с алгеброй чисел определим:

- пусть A — *квадратная невырожденная* (т. е. $\det A \neq 0$) *матрица*; тогда матрица A^{-1} называется *обратной*, если $AA^{-1} = A^{-1}A = I$.

Можно показать, что такое определение обратной матрицы $A^{-1} = (a_{ij}^{\text{обр}})$ приводит к следующей формуле для вычисления ее элементов $a_{ij}^{\text{обр}}$:

$$a_{ij}^{\text{обр}} = \frac{(-1)^{i+j} \det A_{ji}}{\det A}, \quad (\text{П2.19})$$

где A_{ji} — как и прежде, матрица, получающаяся из матрицы A вычеркиванием из нее j -й строки и i -го столбца (т. е. числитель правой части (П2.19) является *алгебраическим дополнением* элемента a_{ji} в *исходной* матрице A , или, что то же, — алгебраическим дополнением (i, j) -го элемента в *транспонированной* матрице A^T).

Пример П2.1. *Двухпродуктовая версия статической модели «затраты-выпуск» Леонтьева.* Пусть a_{ij} — затраты продукта i на выпуск единицы продукта j ($i, j = 1, 2$). И пусть x_i — общий выпуск продукта i и c_i — конечный спрос на этот продукт. Тогда уравнения, связывающие между собой введенные выше величины, будут иметь вид (в предположении, что нет ни потерь, ни излишков продуктов):

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + c_1 = x_1, \\ a_{21}x_1 + a_{22}x_2 + c_2 = x_2, \end{cases}$$

или, после приведения подобных членов,

$$\begin{cases} (1 - a_{11})x_1 - a_{12}x_2 = c_1, \\ -a_{21}x_1 + (1 - a_{22})x_2 = c_2. \end{cases}$$

Запишем эту систему, используя матричные обозначения:

$$AX = C, \quad (\text{П2.20})$$

где $X = (x_1, x_2)^T$, $C = (c_1, c_2)^T$, а

$$A = \begin{pmatrix} 1 - a_{11} & -a_{12} \\ -a_{21} & 1 - a_{22} \end{pmatrix}.$$

Мы хотим разрешить систему (П2.20) относительно x_1 и x_2 , т. е. определить, какое общее количество первого и второго продукта должно быть произведено, чтобы обеспечить производственное потребление и конечный спрос. Чтобы уединить X в уравнении (П2.20), домножим обе части этого уравнения на матрицу A^{-1} слева:

$$X = A^{-1}C,$$

где в соответствии с (П2.19)

$$A^{-1} = \frac{1}{\det A} \begin{pmatrix} 1 - a_{22} & a_{12} \\ a_{21} & 1 - a_{11} \end{pmatrix}.$$

С учетом того, что $\det A = (1 - a_{11})(1 - a_{22}) - a_{12}a_{21}$, имеем:

$$\begin{aligned} x_1 &= \frac{1}{(1 - a_{11})(1 - a_{22}) - a_{12}a_{21}} [(1 - a_{22})c_1 + a_{12}c_2], \\ x_2 &= \frac{1}{(1 - a_{11})(1 - a_{22}) - a_{12}a_{21}} [a_{21}c_1 + (1 - a_{11})c_2]. \end{aligned}$$

Основные свойства обратной матрицы:

1) матрица A^{-1} для любой невырожденной матрицы A — *единственна*;

$$2) \det A^{-1} = (\det A)^{-1};$$

$$3) (A^T)^{-1} = (A^{-1})^T;$$

$$4) (A^{-1})^{-1} = A;$$

$$5) (AB)^{-1} = B^{-1}A^{-1},$$

$$(ABC)^{-1} = C^{-1}B^{-1}A^{-1} \text{ и т. д.}$$

(напомним, что все матрицы, участвовавшие в формулировке свойств обратной матрицы, — квадратные и невырожденные).

Теперь мы можем дополнить перечень основных типов матриц, приведенный в п. П2.1. Введем в рассмотрение *класс ортогональных матриц*, определив:

- *квадратная невырожденная матрица A называется ортогональной, если $A^T = A^{-1}$.*

Из определения ортогональной матрицы непосредственно следует, в частности, что $A^T A = AA^T = I$. Нетрудно также вывести, что *определитель ортогональной матрицы всегда равен по абсолютной величине единице*, т. е. $|\det A| = 1$. Действительно, поскольку $\det A = \det A^T$ (свойство 5) из п. П2.2), а $\det(AB) = \det A \cdot \det B$ (свойство 1) из п. П2.2), то для ортогональной матрицы $\det(AA^T) = \det A \det A^T = (\det A)^2 = \det I = 1$. Отсюда получаем, что $|\det A| = 1$, если A ортогональна.

Мы еще будем обращаться к ортогональным матрицам в пп. П2.5 и П2.6.

П2.4. Ранг матрицы и линейная зависимость ее строк (столбцов)

Наряду с рассмотренными в п. П2.2 двумя основными числовыми характеристиками *квадратной* матрицы — ее *определителем* и ее *следом*, в общем случае, включающем и *прямоугольные* матрицы, существует еще одна очень важная числовая характеристика матрицы — ее *ранг*.

Перед тем, как сформулировать строгое определение этого понятия, рассмотрим понятие *линейной зависимости строк* (столбцов) анализируемой $n \times m$ -матрицы A .

Строки $A_i = (a_{i1}, a_{i2}, \dots, a_{im})$, $i = 1, 2, \dots, n$, матрицы A называются линейно зависимыми, если существуют числа $\lambda_1, \lambda_2, \dots, \lambda_n$, не все равные нулю, и такие, что

$$\lambda_1 A_1 + \lambda_2 A_2 + \dots + \lambda_n A_n = 0_{1..m} \quad (\text{П2.21})$$

(здесь $\mathbf{0}_{1..m}$, в соответствии с принятыми выше обозначениями, — это строка длины m , состоящая из нулей).

Аналогично: столбцы $A_{..j} = (a_{1j}, a_{2j}, \dots, a_{nj})^T$, $j = 1, 2, \dots, m$, матрицы A называются **линейно зависимыми**, если существуют числа $\mu_1, \mu_2, \dots, \mu_m$, не все равные нулю, и такие, что

$$\mu_1 A_{..1} + \mu_2 A_{..2} + \dots + \mu_m A_{..m} = \mathbf{0}_{n..1}. \quad (\text{П2.21}')$$

В противном случае строки (столбцы) называются **линейно независимыми**.

Можно доказать, что **максимальное число линейно независимых строк** $n \times m$ -матрицы A совпадает с **максимальным числом ее линейно независимых столбцов** и, одновременно, — с **максимальным порядком ее не равного нулю минора** (напомним, что минором порядка k матрицы A называется определитель $k \times k$ -матрицы, получающейся из матрицы A вычеркиванием из нее $n - k$ строк и $m - k$ столбцов).

- Ранг $n \times m$ -матрицы A (будем обозначать его **ранг** A) **определяется** как **максимальное число ее линейно независимых столбцов**.

Очевидно, что в силу приведенного выше свойства, ранг матрицы A может быть определен и как **максимальное число ее линейно независимых строк**, и как **максимальный порядок ее отличного от нуля минора**. Кстати, последнее определение часто бывает наиболее удобным с точки зрения возможности *практического вычисления* ранга конкретной матрицы. При этом, определяя ранг матрицы как максимальный порядок ее отличного от нуля минора, подразумевается, что достаточно того, чтобы нашелся *хотя бы один* ненулевой минор порядка k , в то время как все миноры порядка $k + 1$ уже будут равны нулю. Так, например, в 4×3 -матрице

$$A = \begin{pmatrix} 4 & 8 & 2 \\ 2 & 4 & 4 \\ 2 & 4 & -2 \\ 3 & 6 & -3 \end{pmatrix}$$

все четыре минора 3-го порядка равны нулю. Следовательно **ранг** $A < 3$. И хотя большинство миноров 2-го порядка тоже равно нулю, все-таки существуют миноры этого порядка, отличные от нуля. А это значит, что **ранг** $A = 2$.

Заметим, что применительно к матрице X наблюдаемых значений объясняющих переменных (см. (П2.4)) в регрессионном анализе **линейная зависимость столбцов означает линейную зависимость объясняющих переменных**.

Из определения ранга матрицы более или менее непосредственно вытекают следующие его основные свойства:

- 1) $\text{ранг } A \leq \min(n, m)$; если $\text{ранг } A = \min(n, m)$, то говорят, что матрица A — это матрица *полного ранга*;
- 2) $\text{ранг } I_n = n$;
- 3) $\text{ранг}(\text{diag}(a_{11}, a_{22}, \dots, a_{nn})) = k$, где k — число ненулевых элементов в ряду $a_{11}, a_{22}, \dots, a_{nn}$;
- 4) $\text{ранг}(AB) \leq \min\{\text{ранг } A, \text{ранг } B\}$;
- 5) $\text{ранг } B = n$, если B — квадратная *невыврожденная* матрица размерности $n \times n$;
- 6) $\text{ранг}(AB) = \text{ранг } A$, если A — произвольная матрица размерности $n \times m$, а B — *невыврожденная* $m \times m$ -матрица;
- 7) $\text{ранг}(BA) = \text{ранг } A$, если A — произвольная матрица размерности $n \times m$, а B — *невыврожденная* $n \times n$ -матрица;
- 8) $\text{ранг}(B_1 A B_2) = \text{ранг } A$, где A — произвольная матрица порядка $n \times m$, а B_1 и B_2 — невырожденные квадратные матрицы размерностей, соответственно, $(n \times n)$ и $(m \times m)$;
- 9) $\text{ранг } A = \text{ранг } A^T$.

Отметим также один важный для эконометрических приложений (особенно для проблемы идентифицируемости модели) факт, связанный с понятием ранга матрицы. Пусть $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^T$ — набор переменных (среди которых, в целях большей общности, мы допускаем присутствие и переменной, тождественно равной единице), а A — матрица размерности $n \times p$. Система уравнений относительно X может быть записана в виде

$$AX = 0_{n,1}. \quad (\text{П2.22})$$

При решении системы (П2.22) важным моментом является соотношение между числом неизвестных и числом линейно независимых уравнений, содержащихся в системе. Так вот, оказывается, что *число линейно независимых уравнений в системе (П2.22) равно рангу матрицы A .*

П2.5. Характеристические (собственные) числа квадратной матрицы и соответствующие им собственные векторы

Широкий круг задач многомерного статистического анализа и эконометрики сводится (в вычислительном плане) к необходимости анализа и

решения параметрического семейства систем уравнений типа

$$(A - \lambda I_n)X = 0_{n,1}, \quad (\text{П2.23})$$

где A — некоторая $n \times n$ -матрица, $X = (x_1, x_2, \dots, x_n)^T$ — вектор-столбец неизвестных, а λ — некоторый числовой параметр. Для того, чтобы существовало нетривиальное (отличное от нулевого) решение, необходимо, чтобы матрица $A - \lambda I_n$ была вырожденной, т. е. необходимо потребовать, чтобы

$$\det(A - \lambda I_n) = 0. \quad (\text{П2.24})$$

Из правил вычисления определителя матрицы (см. (П2.17)) следует, что левая часть (П2.24) представляет собой алгебраический полином от λ степени n , так что соотношение (П2.24) — это алгебраическое уравнение степени n относительно λ . Само это уравнение, а следовательно, и его корни $\lambda_1, \lambda_2, \dots, \lambda_n$ по построению полностью определяются элементами матрицы A . Приходим к определению:

- **Характеристическими (собственными) числами $n \times n$ -матрицы A называются корни характеристического уравнения (П2.24).**

Беря любое из собственных чисел λ_i и подставляя его в исходное соотношение (П2.23), мы получаем уже конкретную систему уравнений (относительно X) вида

$$(A - \lambda_i I_n)X = 0_{n,1}. \quad (\text{П2.24}')$$

Существование нетривиального решения $X(i)$ этого уравнения обеспечивается равенством нулю определителя $\det(A - \lambda_i I_n)$. Соответственно, приходим к еще одному определению:

- **Вектор-столбец $X(i) = (x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)})^T$, являющийся решением уравнения (П2.24'), называется характеристическим (собственным) вектором матрицы A , соответствующим характеристическому (собственному) числу λ_i .**

А поскольку алгебраическое уравнение степени n имеет n корней (среди которых, вообще говоря, могут быть совпадающие и комплексные), то всякая квадратная $n \times n$ -матрица A имеет n собственных чисел (не обязательно различных) и n соответствующих им собственных векторов.

В дальнейшем, отправляясь от интересов прикладной статистики и эконометрики, мы сосредоточим свое внимание на свойствах характеристических чисел и характеристических векторов только действительных симметрических положительно (или неотрицательно) определенных матриц A (будем для краткости их называть далее, соответственно, спо- и

сно-матрицами)¹.

1) Всякая спо- (сно-) $n \times n$ -матрица имеет n положительных (неотрицательных) действительных характеристических чисел; и наоборот, матрица, все характеристические числа которой положительны (неотрицательны) является спо- (сно-) матрицей.

2) Собственные векторы $X(i)$ и $X(j)$, соответствующие разным собственным числам, всегда *взаимно ортогональны*, т. е.

$$X^T(i)X(j) = 0 \quad \text{при} \quad \lambda_i \neq \lambda_j.$$

Поскольку собственный вектор определяется с точностью до коэффициента пропорциональности, то можно пронормировать собственные векторы $X(1), X(2), \dots, X(n)$ так, что все они будут образовывать *ортонормированную систему*, т. е.

$$X^T(i)X(j) = \begin{cases} 1 & \text{при } i = j, \\ 0 & \text{при } i \neq j \end{cases} \quad \text{для всех } i, j = 1, 2, \dots, n.$$

Заметим, что $n \times n$ -матрица A , составленная из столбцов $X(i)$ ($i = 1, 2, \dots, n$), т. е. матрица

$$X = \begin{pmatrix} X(1) & X(2) & \dots & X(n) \end{pmatrix}$$

будет *ортогональной*, т. е. $X^T X = I_n$.

3) Всякая спо- и сно- матрица A может быть приведена с помощью ортогонального преобразования X к диагональному виду, в котором на диагонали будут стоять характеристические числа, т. е.

$$X^T A X = \text{diag} (\lambda_1, \lambda_2, \dots, \lambda_n).$$

П2.6. Некоторые свойства симметричных положительно (неотрицательно) определенных матриц

На протяжении всего данного пункта мы будем полагать (если специально не оговорено другое), что A — *симметрическая положительно*

¹ Симметрическая $n \times n$ -матрица A называется *положительно (неотрицательно) определенной*, если для любого ненулевого вектора $X = (x_1, x_2, \dots, x_n)^T$ выполняется неравенство $X^T A X > 0$ ($X^T A X \geq 0$). Более подробно о симметрических положительно и неотрицательно определенных матрицах см. в п. П 2.6.

определенная матрица размерности n , т. е. для любого ненулевого вектора $X = (x_1, x_2, \dots, x_n)^T$ выполняется неравенство

$$X^T A X > 0. \quad (\text{П2.25})$$

1) При любой $m \times n$ -матрице B матрица $B^T B$ будет всегда симметрична и неотрицательно определена. Действительно, для любой $(n \times 1)$ -матрицы X (т. е. для вектора-столбца длины n)

$$X^T (B^T B) X = (B X)^T B X = Y^T Y = \sum_{i=1}^n y_i^2 \geq 0,$$

т. к. произведение вектора-строки $Y^T = (B X)^T$ на вектор-столбец $Y = B X$ приводит, очевидно, к неотрицательному числу.

2) Если $n \geq m$ и $n \times m$ -матрица B — матрица полного ранга (т. е. *ранг* $B = m$), то матрица $B^T A B$ — положительно определенная; в частном случае $n = m$ B является невырожденной квадратной матрицей размерности $n \times n$.

3) Матрица A^{-1} , обратная к A , также является симметрической и положительно определенной.

4) Единичная матрица I_n является симметрической положительно определенной (т. к. для любого вектора-столбца X длины n $X^T I_n X = X^T X = \sum_{i=1}^n x_i^2 > 0$).

5) $\det A > 0$, т. е. определитель любой споматрицы положителен; отсюда, в частности, следует, что и все *главные миноры* матрицы A положительны (минор матрицы называется *главным*, если он является определителем подматрицы, образованной из исходной матрицы вычеркиванием из нее строк и столбцов, имеющих *одинаковые* номера).

6) След матрицы A может быть вычислен как сумма ее собственных значений, т. е.

$$\text{tr} A = \lambda_1 + \lambda_2 + \dots + \lambda_n.$$

где, напомним, A — споматрица размерности $n \times n$, а $\lambda_1, \lambda_2, \dots, \lambda_n$ — ее собственные числа.

7) Всякая споматрица $(m \times m)$ -матрица A может быть представлена в виде

$$A = C C^T, \quad (\text{П2.25}^a)$$

где C некоторая невырожденная $(m \times m)$ -матрица.

Действительно, из доказанного в п. 13.2.3 соотношения (13.10) следует, что любая спм матрица A может быть приведена к диагональному виду при помощи некоторого ортогонального преобразования L , т. е.

$$LAL^T = \Lambda, \quad (\text{П2.25}^6)$$

где,

$$\Lambda = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_m \end{pmatrix},$$

а $\lambda_1, \lambda_2, \dots, \lambda_m$ — собственные значения (*действительные и положительные*) матрицы A . Домножая обе части соотношения (П2.25⁶) слева на L^T и справа на L , получаем

$$A = L^T \Lambda L. \quad (\text{П2.25}^8)$$

Представляя матрицу A в виде $\Lambda = \Lambda^{1/2} \Lambda^{1/2}$, где

$$\Lambda^{1/2} = \begin{pmatrix} \sqrt{\lambda_1} & & & 0 \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ 0 & & & \sqrt{\lambda_m} \end{pmatrix},$$

и обозначая $C = L^T \Lambda^{1/2}$, получаем из (П2.25⁸) доказательство возможности представления матрицы A в виде (П2.25⁸).

Пример П2.2. Продемонстрируем использование свойств симметрических положительно определенных матриц при анализе многомерных распределений и их статистик. Пусть $X = (x_1, x_2, \dots, x_n)^T$ — последовательность одинаково $(0, \sigma^2)$ -нормально распределенных случайных величин, т. е.

$$Ex_i = 0, \quad i = 1, 2, \dots, n;$$

$$Dx_i = Ex_i^2 = \sigma^2, \quad i = 1, 2, \dots, n;$$

$$\text{cov}(x_i, x_j) = E(x_i x_j) = 0 \quad \text{при} \quad i \neq j,$$

или, в компактной записи, $X \in N_n(0_{n,1}; \sigma^2 I_n)$. И пусть в процессе анализа модели нам пришлось заниматься исследованием квадратичной формы от анализируемых случайных величин вида $X^T A X$, где A , как и ранее, спм-матрица.

В соответствии со свойством 3) спм- и сно-матриц (см. п. П2.5) существует такая ортогональная матрица C , с помощью которой матрица

A может быть приведена к диагональному виду, а именно:

$$C^T A C = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n),$$

где $\lambda_1, \lambda_2, \dots, \lambda_n$ — собственные числа матрицы A , а элементы матрицы C полностью определяются собственными векторами той же матрицы A . Введем вспомогательные переменные

$$Y = C^T X \quad (\text{соответственно, } X = CY) \quad (\text{П2.26})$$

и посмотрим, как будет выражаться анализируемая квадратичная форма $X^T A X$ в терминах этих новых переменных:

$$\begin{aligned} X^T A X &= (CY)^T A CY = Y^T (C^T A C) Y \\ &= Y^T \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) Y = \sum_{i=1}^n \lambda_i y_i^2. \end{aligned} \quad (\text{П2.27})$$

Следовательно, для того чтобы понять, как «ведет себя» квадратичная форма $X^T A X$, необходимо выяснить, какому закону распределения вероятностей подчиняются переменные $y_1, y_2, \dots, y_n$.

Из (П2.26) следует, что Y , будучи линейной функцией от X , также подчиняется нормальному закону со средним значением

$$EY = E(C^T X) = C^T EX = 0_{n,1}$$

и с ковариационной матрицей

$$\begin{aligned} \Sigma_Y &= E(YY^T) = E(C^T X X^T C) = C^T E(X X^T) C \\ &= C^T \Sigma_X C = C^T (\sigma^2 I_n) C = \sigma^2 C^T C = \sigma^2 I_n \end{aligned}$$

(в этой выкладке мы воспользовались тем, что ковариационная матрица Σ_X вектора X равна по условию $\sigma^2 I_n$, а также — ортогональностью матрицы C , из чего следует $C^T C = I_n$).

Таким образом, мы убедились в том, что вспомогательные переменные $y_1, y_2, \dots, y_n$ так же, как и исходные случайные величины $x_1, x_2, \dots, x_n$, $(0, \sigma^2)$ -нормальны и взаимно независимы. А это позволяет описать распределение интересующей нас квадратичной формы (П2.27) законом суммы масштабированных $\chi^2(1)$ — распределенных случайных величин.

П2.7. Идемпотентные матрицы

Особенно важны в прикладной статистике и эконометрике соотношения типа (П2.27), в которых матрица A относится к классу так называемых *идемпотентных матриц*.

- **Идемпотентной** называется симметрическая¹ матрица A , обладающая тем свойством, что она совпадает со своим квадратом, т. е.

$$A = A^2. \quad (\text{П2.28})$$

Перечислим основные свойства идемпотентных матриц.

1) Любая целая положительная степень k идемпотентной матрицы A дает снова исходную матрицу A , т. е. $A^k = A$ (следует непосредственно из определения).

2) Собственные числа идемпотентной матрицы могут принимать только одно из двух значений: 0 или 1; а следовательно, в соответствии со свойством 1) из п. П2.5 все идемпотентные матрицы неотрицательно определены.

Докажем справедливость этого утверждения. В соответствии с характеристическим уравнением (П2.23)

$$A X = \lambda X.$$

После домножения обеих частей этого характеристического уравнения слева на матрицу A имеем:

$$A^2 X = \lambda (A X) = \lambda (\lambda X) = \lambda^2 X. \quad (\text{П2.29})$$

Но одновременно с этим, используя идемпотентность матрицы A , мы можем записать:

$$A^2 X = A X = \lambda X. \quad (\text{П2.29}')$$

Сравнивая правые части (П2.29) и (П2.29'), получаем

$$(\lambda^2 - \lambda)X = 0_{n,1},$$

а т. к. $X \neq 0_{n,1}$, то

$$\lambda^2 - \lambda = 0,$$

т. е. либо $\lambda = 0$, либо $\lambda = 1$, что и доказывает свойство 2).

¹ Иногда требование симметричности не включают в определение идемпотентной матрицы.

3) Ранг идемпотентной матрицы равен числу ее ненулевых собственных чисел, что совпадает со следом исходной матрицы.

4) Если ранг $n \times n$ -идемпотентной матрицы A равен $n - k$, то использование ее собственных векторов для преобразований типа (П2.26) будет приводить квадратичную форму $X^T A X$ относительно n независимых $(0, \sigma^2)$ -нормально распределенных случайных величин к виду

$$X^T A X = y_1^2 + y_2^2 + \dots + y_{n-k}^2, \quad (\text{П2.30})$$

где элементы $y_1, y_2, \dots, y_{n-k}$ также взаимно независимы и $(0, \sigma^2)$ — нормально распределены. А это значит, что квадратичная форма $X^T A X$ подчиняется $\sigma \chi^2(n - k)$ -распределению.

П2.8. Матрицы блочной структуры

Содержание анализируемой задачи зачастую обуславливает целесообразность разбиения матриц, которыми мы оперируем, на отдельные блоки. Так, например, при анализе модели многомерной регрессии или систем одновременных уравнений в рамки нашего анализа попадают m зависимых (эндогенных) переменных $Y = (y^{(1)}, y^{(2)}, \dots, y^{(m)})^T$ и p объясняющих (экзогенных, предопределенных) переменных $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})^T$. Важным объектом анализа является при этом *ковариационная матрица Σ_Z всего вектора переменных $Z = \begin{pmatrix} Y \\ X \end{pmatrix}$* , которая будет состоять, соответственно, из четырех блоков Σ_{YY} , Σ_{YX} , Σ_{XY} и Σ_{XX} , а именно:

$$\Sigma_Z = \begin{pmatrix} \Sigma_{YY} & \vdots & \Sigma_{YX} \\ \dots\dots\dots & & \\ \Sigma_{XY} & \vdots & \Sigma_{XX} \end{pmatrix}. \quad (\text{П2.31})$$

Очевидно, матрица Σ_Z будет иметь размерность $(m+p) \times (m+p)$, в то время как блок Σ_{YY} (матрица ковариаций между компонентами вектора Y) имеет размерность $m \times m$, блок Σ_{YX} (матрица «перекрестных» ковариаций эндогенных переменных с предопределенными) имеет размерность $m \times p$, блок Σ_{XY} (матрица «перекрестных» ковариаций предопределенных переменных с эндогенными) имеет размерность $p \times m$, и, наконец, блок Σ_{XX} (матрица ковариаций между компонентами вектора X) имеет размерность $p \times p$.

Заметим, что для того, чтобы пользоваться некоторыми общими правилами оперирования с блочными матрицами, следует соблюдать *необходимые* требования при их разбиении, а именно:

- блоки, стоящие друг *над* (или *под*) другом *по вертикали*, должны иметь *одинаковое количество столбцов*; а блоки, стоящие *рядом* (слева или справа друг от друга) *по горизонтали*, должны иметь *одинаковое количество строк*.

Сложение блочных матриц. Если две матрицы A и B одинаковой размерности разбиты на блоки, соответственно, A_{ij} и B_{ij} ($i = 1, 2, \dots, k_1$; $j = 1, 2, \dots, k_2$) *одинаковым образом* (т. е. размерности блоков A_{ij} и B_{ij} совпадают), то их сложение производится по правилу

$$A + B = \begin{pmatrix} A_{11} + B_{11} & \vdots & A_{12} + B_{12} & \vdots & \dots & \vdots & A_{1k_2} + B_{1k_2} \\ A_{21} + B_{21} & \vdots & A_{22} + B_{22} & \vdots & \dots & \vdots & A_{2k_2} + B_{2k_2} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ A_{k_11} + B_{k_11} & \vdots & A_{k_12} + B_{k_12} & \vdots & \dots & \vdots & A_{k_1k_2} + B_{k_1k_2} \end{pmatrix}.$$

Умножение блочных матриц формально также производится по правилам умножения *обычных* матриц, однако разбиение A и B на блоки A_{ij} и B_{ij} должно быть произведено таким образом, чтобы количество столбцов в первом сомножителе совпадало бы с количеством строк во втором сомножителе. Например:

$$\begin{aligned} AB &= \begin{pmatrix} A_{11} & \vdots & A_{12} \\ \dots & \dots & \dots \\ A_{21} & \vdots & A_{22} \end{pmatrix} \begin{pmatrix} B_{11} & \vdots & B_{12} \\ \dots & \dots & \dots \\ B_{21} & \vdots & B_{22} \end{pmatrix} \\ &= \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} & \vdots & A_{11}B_{12} + A_{12}B_{22} \\ \dots & \dots & \dots \\ A_{21}B_{11} + A_{22}B_{21} & \vdots & A_{21}B_{12} + A_{22}B_{22} \end{pmatrix}. \end{aligned}$$

Определитель блочной матрицы. Пусть матрица A разбита на 4 блока

$$A = \begin{pmatrix} A_{11} & \vdots & A_{12} \\ \dots & \dots & \dots \\ A_{21} & \vdots & A_{22} \end{pmatrix} \quad (\text{П2.32})$$

таким образом, что блоки A_{11} и A_{22} являются *квадратными* матрицами (как это было в нашем примере (П2.31)). Тогда можно воспользоваться следующей общей формулой подсчета определителя матрицы A , во многих

случаях существенно упрощающей вычисления:

$$\det \begin{pmatrix} A_{11} & \vdots & A_{12} \\ \dots\dots\dots & & \\ A_{21} & \vdots & A_{22} \end{pmatrix} = \det A_{11} \cdot \det (A_{22} - A_{21} A_{11}^{-1} A_{12}) \\ = \det A_{22} \cdot \det (A_{11} - A_{12} A_{22}^{-1} A_{21}).$$

В частном случае *блочно-диагональной* матрицы A (т. е. когда матрицы A_{12} и A_{21} состоят из одних нулей) данная формула принимает вид

$$\det \begin{pmatrix} A_{11} & \vdots & 0 \\ \dots\dots\dots & & \\ 0 & \vdots & A_{22} \end{pmatrix} = \det A_{11} \cdot \det A_{22}.$$

Обращение блочной матрицы (П2.32) производится по формуле

$$\begin{pmatrix} A_{11} & \vdots & A_{12} \\ \dots\dots\dots & & \\ A_{21} & \vdots & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} A^{11} & \vdots & -A_{11}^{-1} A_{12} A^{22} \\ \dots\dots\dots & & \\ -A^{22} A_{21} A_{11}^{-1} & \vdots & A^{22} \end{pmatrix},$$

где

$$A^{22} = (A_{22} - A_{21} A_{11}^{-1} A_{12})^{-1} \quad \text{и} \quad A^{11} = (A_{11} - A_{12} A_{22}^{-1} A_{21})^{-1}.$$

Прямое (кронекерово) произведение $n \times m$ -матрицы A и $k \times l$ -матрицы B было введено в п. П2.1 (см. соотношение (П2.15)). Поскольку результат такого перемножения матриц A и B представляется *матрицей блочной структуры*, рассмотрим несколько полезных свойств прямого произведения:

1) если матрицы A и B квадратны и обратимы, то

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1};$$

2) если A — $n \times n$ -матрица, а B — $k \times k$ -матрица, то

$$\det (A \otimes B) = (\det A)^n (\det B)^k;$$

$$3) (A \otimes B)^T = (A^T \otimes B^T);$$

$$4) \operatorname{tr} (A \otimes B) = \operatorname{tr} A \cdot \operatorname{tr} B.$$

ЛИТЕРАТУРА

- Абрамовиц М.* Справочник по специальным функциям. /Пер. с англ./ — М.: Наука, 1979.
- Айвазян С. А.* (1968). Статистическое исследование зависимостей. М.: Металлургия.
- Айвазян С. А.* (1974). Об опыте применения экспертно-статистического метода построения неизвестной целевой функции. — В кн.: Многомерный статистический анализ в социально-экономических исследованиях. М.: Наука, С. 56–86.
- Айвазян С. А.* (1979). Конфлюэнтный анализ. — В кн.: Математическая энциклопедия. М. Т. 2. С. 1083.
- Айвазян С. А.* (1980). Вероятностно-статистическое моделирование распределительных отношений в обществе. — В кн.: Статистическое моделирование экономических процессов. Ученые записки по статистике. М., Статистика, Т.37.
- Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика. Исследование зависимостей. — М.: Финансы и статистика, 1985.
- Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика. Классификация и снижение размерности. — М.: Финансы и статистика, 1989.
- Андерсон Т.* Введение в многомерный статистический анализ. М., Физматгиз, 1963.
- Бокс Дж., Дженкинс Г.* Анализ временных рядов. Прогноз и управление. /Пер. с англ./ — М.: Мир, 1974, вып. 1 и 2.
- Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. М.: Наука, 1965.
- Боровков А. А.* Курс теории вероятностей. М.: Наука, 1972.
- Бухштабер В. М.* Метод многомерной развертки в анализе временных рядов. Обзор прикладной и промышленной математики. Т.4 (1997): № 4.
- Вальд А.* Последовательный анализ. /Пер. с англ./ — М.: Физматгиз, 1960.
- Джонстон Дж.* Эконометрические методы. /Пер. с англ./ — М.: Статистика, 1980.
- Дидэ Э., и др.* Методы анализа данных: подход, основанный на методе динамических сгущений. /Пер. с франц./ — М.: Финансы и статистика, 1985.
- Езекиэл М., Фокс К.* Методы анализа корреляций и регрессий. /Пер. с англ./ — М.: Статистика, 1966.
- Кемени Дж., Снелл Дж.* Конечные цепи Маркова. /Пер. с англ./ — М.: Наука, 1970.
- Кендэл М.* Временные ряды. /Пер. с англ./ — М.: Финансы и статистика, 1981.
- Кендалл М. Дж., Стьюарт А.* (1966). Теория распределений. /Пер. с англ./ — М.: Наука.
- Кендалл М. Дж., Стьюарт А.* (1973). Статистические выводы и связи. /Пер. с англ./ — М.: Наука.
- Кендалл М. Дж., Стьюарт А.* (1976). Многомерный статистический анализ и временные ряды. /Пер. с англ./ — М.: Наука.
- Колмогоров А. Н.* Основные понятия теории вероятностей. М. — Л., ОНТИ, 1936.

Крамер Г. Математические методы статистики. — 2-е изд. Пер. с англ. — М.: Мир, 1975.

Лумельский В. Я. Агрегирование объектов на основе квадратичной матрицы. Автоматика и телемеханика, 1970. № 1. С.133–143.

Магнус Я. Р., Катышев П. К., Персецкий А. А. Эконометрика. Начальный курс. (3-е изд.) — М.: Дело, 2000

Майстров Л. Е. Развитие понятия вероятности. — М.: Наука, 1980.

Маленко Э. Статистические методы в эконометрии: Пер. с франц. — М.: Статистика, 1975, вып.1; 1976, вып.2.

Митоян А. А. Потребительское поведение семей: дифференциация, динамика, классификация. — М.: Экономика, 1990.

Феллер В. Введение в теорию вероятностей и ее приложения: Пер. с англ. — 2-е издание. М.: Мир, 1964.

Харман Г. Современный факторный анализ: Пер. с англ. — М.: Статистика, 1972. Мир, 1983.

Эфрон Б. Нетрадиционные методы многомерного статистического анализа: Пер. с англ. — М.: Финансы и статистика, 1988.

Anderson T. W., Rubin H. Statistical inferences in factor analysis. Proc. 3 Berkeley Symp. Math. Statist. and Probab. — Univ. Calif. Press, 1956. Pp. 11–50.

Berndt E. R. The practice of econometrics. Classic and contemporary. Addison-Wesley Publishing Company. Reading–Massachusetts–Menlo Parc–California, 1990.

Dougherty C. Introduction to econometrics. Oxford University Press. New York–Oxford, 1992.

Goldberger A. A course in Econometrics. Cambridge–Mass.: Harvard University Press, 1990.

Green W. H. Econometric analysis. Macmillan Publishing Company, New York, 1993.

Hayashi F. Econometrics. — Princeton University Press, Princeton and Oxford, 2000.

Horowitz J. Semiparametric Methods in Econometrics. — Springer-Verlag New York, Inc, 1998.

Johnston J. and Di Nardo J. Econometric Methods, 4th edition. — Mc Graw-Hill, 1997.

Jorgenson D. Welfare. Vol.1: Aggregate Consumer Behavior. Vol.2: Measuring Social Welfare. — MIT Press, Cambridge, Massachusetts, 1997.

Kennedy P. A Guide to Econometrics, 4th edition. — Blackwell Publishers, 1998.

Magnus J. R., Neudecker H. Matrix Differential Calculus with Applications in Statistics and Econometrics. New York, John Wiley, 1988.

Pindyck R., Rubinfeld D. L. Econometric models and economic forecasts. McGraw-Hill Kogakusha Ltd, Tokio, 1976.

Ruud P. An Introduction to Classical Econometric Theory. — Oxford University Press. New York–Oxford, 2000.

АЛФАВИТНО-ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

А

Автоинформативность 336, 524–525
Автоматическая классификация 480, 487
Автопрогноз 541
Аксиоматика Колмогорова 69–70
Алгебраическое дополнение 413, 627
Алгебраический полином 383
Апостериорное распределение 271
Апостериорно-модельный подход 62
Апостериорно-частотный подход 61
Апостериорный байесовский риск 272
Априорная статистическая (выборочная) информация 469
Априорное распределение 270
Априорный подход к вычислению вероятностей 60
Асимметрия 108, 210
Асимптотическая несмещенность 251
— эффективность 251
Асимптотические свойства оценок 234, 251

Асимптотическое распределение 216

Б

Байеса теорема (формула) 69, 271–272
Байесовские оценки параметров 271–272
Байесовский подход 269
Байесовский риск 272
Байесовское правило классификации 473, 474
Безусловная вероятность 177
Бернулли последовательность испытаний 115
— теорема 155
Бета-распределение 148, 276
Бинарные (булевы, дихотомические) переменные 493
Блочно-диагональная матрица 641
Блочная структура матриц 639–641
Булева матрица парных сравнений

584

Бюджетные обследования семей 353

В

Вариационный ряд 225

Вероятностная модель 41, 50

Вероятностное пространство 59, 195

— — дискретное 52

— — непрерывное 69

Вероятностно-статистическая модель
40–41, 50Вероятность ошибочной классифика-
ции 473

Вероятность события 59–60

— — условная 64

Взвешенное евклидово расстояние
493Взвешивание выборочных данных
261

Визуализация данных 336

Возможные значения случайной ве-
личины 79Восстановление пропущенных наб-
людений 342Выбор общего вида модели
382–385

Выборка 195, 329

— механическая 199

— случайная 199

— ступенчатая 200

Выборочные характеристики 209–
210

Г

Гамма-распределение 146

Гамма-функция Эйлера 142, 145, 147

Генеральная совокупность 194, 329

Геометрическая природа совокупно-
сти исходных данных 339, 383, 394Гипергеометрическое распределение
119–121

Гипотеза статистическая 283–287

— — простая 289, 292

— — сложная 290, 293

— — нулевая (основная) 287

— — альтернативная 287

Гистограмма 203, 462

Главные компоненты 526–551

Группированные выборочные дан-
ные 207

Д

Двумерный нормальный закон рас-
пределения 97, 127

Детерминант (опредетель) 626

Дендограмма 480, 504

Дескриптивная статистика 231

Дискриминантная функция 469, 474

Дискриминантный анализ 335

— — параметрический 475, 476

— — непараметрический 476

Дисперсионный анализ 314, 363, 393

Дисперсия 104

- обобщенная 110
- эмпирическая 209
- Дихотомические (бинарные) переменные 493
- Доверительная вероятность 33, 248
- область 249, 264
- Доверительный интервал 248, 280
- Достоверное событие 57

З

- Зависимость детерминированная 371
- марковского типа 169
- статистическая 359
- структурного типа 374
- Закон больших чисел 154–155
- Закон распределения вероятностей 84, 150
- — — многомерный 86
- Зона (область) неопределенности 299

И

- Иерархические процедуры классификации 497, 504–507
- Интервал группирования 386
- Интервальное оценивание 248
- Информационная матрица Фишера 243
- характеристика связи 460
- расстояние Каллбэка 498

- Информационный этап исследования 340
- Испытания Бернулли 171
- Исходные статистические данные 43, 329–333, 447
- Итоговый этап исследования 345

К

- Категоризованные переменные 448
- Квантиль 106
- Класс 471
- Классификация 459
- без обучения 460
- с обучением 460
- Классификационные переменные 362
- Кластер 334, 489
- Кластер-анализ 334, 480, 491
- Ковариационная матрица 101, 110, 397
- Ковариационный анализ 363
- Количественные переменные 362
- Количество информации (Фишера) 242
- Комбинационная группировка 458, 462
- Комбинированный (ступенчатый) отбор 200
- Конфлюентный анализ 374
- Корреляционная матрица 111, 533
- Корреляционное отношение 412
- поле 385
- Корреляционный анализ 377

- асимметрии 108
- вариации 105
- детерминации 399, 401, 403, 424, 425, 454
- дифференциации 107
- квадратической сопряженности 449
- конкордации (согласованности) 442–446
- корреляции 111, 397, 404, 405
- — выборочный 388, 407
- — множественный 422–425
- — парный 111, 404, 405
- — ранговый (см. «Ранговый коэффициент корреляции»)
- — частный 418, 419
- эксцесса 108

Критерии статистические 283, 290–294

- — наиболее мощные 291, 293

Критерий автоинформативности 524–525

- Барлетта 316
- Вальда 299–300
- Вилкоксона–Манна–Уитни 311
- внешней информативности 524
- дисперсионного анализа 314, 315
- информативности 335, 523, 532
- качества аппроксимации (адекватности модели) 351
- Колмогорова–Смирнова 308
- логарифма отношения правдоподобия 292
- — наиболее мощный 291–292

- несмещенный 295
- однородности дисперсий 315
- отношения правдоподобия 291, 318
- последовательный 299–300, 325
- равномерно наиболее мощный 292
- Смирнова 309
- согласия 284, 300–309
- состоятельный 295
- Стьюдента (t -критерий) 313

Критическая статистика 288

Кумулянта (кумулятивная кривая) 205

Л

Латентно-структурный анализ 337

Латентный показатель 335, 580, 581

Линейные комбинации нормально распределенных величин 219

Логарифмическая функция правдоподобия 252

Логарифмически-нормальный закон распределения 129

Логико-алгебраический подход 330

М

Марковская цепь 169

Массив исходных статистических данных 331, 447–448

Математико-статистические методы
39

Математическая модель 40

Математическая статистика 29, 48

Математическое ожидание 99

Матрица 619

— квадратная 620

— ковариационная 101, 110, 397, 533

— корреляционная 111, 533

— идемпотентная 637

— нагрузок 537

— невырожденная 627

— неотрицательно определенная 634

— обратная 628

— «объект-свойство» 331

— парных сравнений 332

— переходных вероятностей 170

— полного ранга 635

— положительно определенная 634

— симметрическая 620

— транспонированная 621

Мера близости 491

Медиана 103

— главных компонент, см. Главные
компоненты

— корреляционных плеяд 577

— k -средних 513

— максимального правдоподобия
249–258

— моментов 258

Методы машинного ассистирования
340

Многомерная случайная величина

80–81

— функция плотности 95

— — распределения 94

Многомерное наблюдение 348

— шкалирование 337, 592

— — метрическое 592–594

— — неметрическое 595

Многомерный статистический ана-
лиз 362

Множественная линейная регрессия
371

Множественный коэффициент корреле-
ляции 422–425

Мода 103

— вероятностная 40–41, 50

— вероятностно-статистическая 40–
41, 50

— закона распределения 114

— многомерной развертки времен-
ного ряда 339

— регрессионная 359, 371

— смеси распределений 473, 485

— факторного анализа 551–553

— эконометрическая 41

— экономическая 41

Мощность статистического крите-
рия 289, 294

Мультиколлинеарность 378

Н

Накопленная частота 90

Начальное распределение вероятно-

стей 169

Начальные моменты случайной величины 100

— — выборочные 207

Невозможные события 57

Независимость наблюдений 196–197

— случайных величин 98

— событий 66

Незамкнутый граф 574

Неидентифицируемость 558

Неймана–Пирсона лемма 291

Непараметрические (или частично-параметрические) методы 378

Непараметрический дискриминантный анализ 476

Непрерывное вероятностное пространство 69

Неравенство информации (Рао–Крамера–Фреше) 249

— Чебышева 153

Неразличимые («связные», «объединенные») ранги 429, 436

Несмещенность оценки 236

Несовместные события 57

Нормальный (гауссовский) закон распределения 125

О

Обобщенная дисперсия 110

Обобщенный коэффициент корреляции 438

Обращение матрицы 479, 628

— — блочной 640

Обучающая выборка 459, 469, 480

Общие факторы (в факторном анализе) 553

Объем выборки 197

Объясняющая переменная 351

Обычное евклидово расстояние 493

Однородная цепь Маркова 169

Однородность объектов 491

Оперативная характеристика критерия 294

Описательная (дескриптивная) статистика 231

Определитель (детерминант) блочной матрицы 640

— — матрицы 626

Оптимальная (байесовская) процедура классификации 474

Оптимизационные (тремальные) формулировки статистических задач 330

Основные выборочные характеристики 201

— типы зависимостей 373

Отбор информативных типобразующих признаков 335, 463

Отклонение среднеквадратическое 105, 209

Отношение правдоподобия 290–291

— предпочтения 587

Отрицательный биномиальный закон 118

Оценка значений общих факторов 559

— — — — метод Бартлетта 564

- — — — модель Томсона 565
- максимального правдоподобия 250, 227
- параметров уравнения статистической связи 379
- статистическая 233
- — несмещенная 236
- — состоятельная 234
- — эффективная 239
- Ошибка аппроксимации 400
- второго рода 289
- выборки 198
- классификации 471
- первого рода 289
- Оцифровка 344

П

- Параллельные кластер-процедуры 507–508
- Параметризация генеральной совокупности 344
 - многомерного распределения 109–111
 - регрессионной модели 394
- Параметрические регрессионные схемы 382
- Параметрический дискриминантный анализ 475–477, 487
- Парные сравнения 587, 592
- Пассивный эксперимент 341
- Первичная статистическая обработка данных 341, 461
- Переходные вероятности 172
- План вычислительного анализа данных 345
- План сбора исходной статистической информации 340
- Плотность вероятности 92
- Подведение итогов исследования 345
- Показатели качественные (порядковые, ординальные) 343
 - классификационные (номинальные) 343
 - количественные 343
- Полигон 205
 - распределения (вероятностей) 85
- Полиномиальная регрессия 383
- Полиномиальное (мультиномиальное) распределение 123, 226
- Полная система событий 57
- Последовательная схема наблюдений 297
- Последовательные кластер-процедуры 5049–505
- Последовательный критерий Вальда 299
- Предикторы в модели регрессии 336
- Преобразование Бокса–Кокса 392
- Прикладная статистика 27, 49, 346
- Принцип отношения правдоподобия 288, 290
- Проблема группового выбора (упорядочения) 432
- Проверка однородности исходных данных 342
 - статистической независимости

последовательности наблюдений 343

— значений результирующего показателя 365–366

Произведение (пересечение) событий 56

Пропущенные (стертые) наблюдения 342

Простая типологизация 465

Простой случайный отбор 199

Пространственная выборка 331

Пространственно-временная выборка 331

Пространство элементарных событий 53

Противоположное (дополнительное) событие 56

Процедуры классификации 469, 474

— — иерархические 504, 505–507

— — параллельные 504, 507–509

— — последовательные 504

Процентная точка распределения 106

Прямое (кронекерово) произведение матриц 640

Р

Равномерно наиболее мощный критерий 292

Разность событий 56

Ранг матрицы 630–632

— объекта 428–429

Ранговая корреляция 379, 381

Ранговый коэффициент корреляции

349, 430

— — — Кендалла 384, 434

— — — — случай связных рангов 436

— — — Спирмэна 383, 432

— — — — случай связных рангов 433

Распределения вероятностей закон 84, 150

— — — бета (β) 148

— — — биномиальный 116

— — — Вейбулла 136

— — — гамма (γ) 141, 146

— — — гипергеометрический 119

— — — двусторонний экспоненциальный (Лапласа) 138

— — — Коши 140

— — — логарифмически-нормальный 129

— — — нормальный (гауссовский) 125

— — — — двумерный 127

— — — — многомерный 129

— — — отрицательный биномиальный 118

— — — Парето 139

— — — полиномиальный (мультиномиальный) 123

— — — Пуассона 121

— — — равномерный (прямоугольный) 132

— — — совместный (многомерный) 86, 94–95

— — — Стьюдента (t) 143

— — — условный 88, 96

— — — F (Фишера) 144
 — — — «хи-квадрат» (χ^2) 141
 — — — частный (маргинальный)
 87, 96

— — — экспоненциальный (показательный) 134, 137

Рассеяние внутриклассовое 496–497

Расслоенная выборка 200

Расстояние между классами объектов 495

— — — — «ближнего соседа» 495

— — — — «дальнего соседа» 495

— — — — информационное (Каллбэка) 498

— — — — Махаланобиса 498

— — — — обобщенное (по Колмогорову) 496–497

— — — — «средний связи» 496

— — — — «центров тяжести» 495

— — объектами 491

— — — взвешенное евклидово 493

— — — евклидово 492

— — — хеммингово 493

Расщепление смеси распределений 335, 486

Реализация временного ряда 331

Реальный комплекс условий (случайного эксперимента) 30

Регрессионная зависимость 371

Регрессионный анализ 352

Регрессия 359

Результирующая переменная 351

Репрезентативность выборки 198

Решающее правило 474

Риск байесовский апостериорный 272

Робастные (устойчивые) оценки 252

Ряд вариационный, см. Вариационный ряд

С

Сбор статистических данных 340

Сводка и группировка данных 341

Серийная выборка 200

Симметрическая матрица 620

Сжатие массивов информации 336

Система одновременных уравнений 364, 370

Систематизация прикладных задач классификации 465–468

Ситуационный анализ 370

Случайная величина 78, 112

— — дискретная 80

— — многомерная (векторная) 81

— — некатегоризованная 83

— — непрерывная 80

— — номинальная (классификационная) 82

— — одномерная (скалярная) 81

— — ординальная (порядковая) 82

След матрицы 627

Случкогo теорема 215

Случайное событие 53, 55, 71

— — достоверное 57

— — невозможное 57

- — противоположное 56
- — элементарное 53
- Смесь распределений вероятностей 480, 482
- двух нормальных законов распределения 481
- многомерных нормальных законов распределения 484
- Снижение размерности исследуемого пространства 520
- Сно-матрицы 634–637
- Собственное (характеристическое) значение (число) матрицы 531, 633
- Собственный (характеристический) вектор матрицы 531, 533, 633
- Совместное (многомерное) распределение см. Распределения вероятностей закон совместный
- Состояние поглощающее (замкнутое) 178, 179
- возвратное 179
- достижимое 178
- периодическое 179
- эргодическое 179
- Состоятельность оценки 234
- Сплайн 378
- Спо-матрицы 634–637
- Способы организации выборки 198–200
- Среднее значение выборочное 208
- — геометрическое 102, 208
- — гармоническое 103, 208
- — теоретическое 99
- Среднеквадратическое отклонение 105
- Средние удельные потери 472
- Средний объем выборки 297
- Средняя мера внутриклассового рассеяния 501–502
- Стандартная нормальная функция распределения 213
- Статистика 233
- Статистика предметно-содержательная (социально-экономическая и др.) 27–28
- Статистическая гипотеза 283
- — об однородности нескольких выборок 285
- — о виде модели статистической зависимости 286
- — о типе закона распределения 284
- — о числовых значениях параметров 285–286
- Статистическая независимость случайных величин 98
- Статистическая проверка гипотез 283
- Статистическая оценка 233
- Статистически значимая связь 450
- Статистические методы классификации 333
- Статистическое исследование зависимостей 333, 349, 351
- моделирование 133
- оценивание параметров 232
- Степень согласованности мнений группы экспертов 436
- тесноты статистической связи 352
- Стратифицированный (расслоенный) отбор 199

Структура модели 353, 361
Сумма (объединение) событий 55
Сходимость по вероятности 154

Т

Таблица (матрица) «объект-свойство» 331

Таблица сопряженности 396, 448

Теорема (формула) Байеса 69, 271–272

— Бернулли 155

— Вейерштрасса (см. Вейерштрассе теорема)

— Муавра–Лапласа 159

— сложения вероятностей 63

— Слуцкого 215

— умножения вероятностей 64

Теоретико-вероятностная модель 40

Теория вероятностей 28, 39, 48

Типологизация задач снижения размерности 524

— математических постановок задач классификации 469

— простая 465

— связная неупорядоченная 466

— — упорядоченная 467

— структурная 467

Типообразующие факторы (признаки) 335, 461

Томография 339

Точечная оценка 248

Транспонирование матрицы 621

— произведения матриц 623

У

Унификация типа переменных 343

— характеристическое матрицы 633

Урезание выборки 261–262

— распределения 263

Уровень значимости критерия 33, 286

Условная вероятность 64

— плотность вероятности 96, 128

Условный закон распределения вероятностей 88, 96–97

Устойчивость статистическая (выборочных характеристик) 214–215

Устойчивые статистические выводы 339

Ф

Факторные нагрузки 553, 559

Факторный анализ 551–553

Формула Байеса 68, 271–272

— композиции (свертки) 164

— полной вероятности 67

Функционал (критерий) качества метода или модели 329, 351, 357, 498–503, 519, 521–523, 527, 555–556, 573

Функциональная зависимость 373, 402

Функция мощности критерия 294

Функция плотности вероятности 92, 112, 125, 127, 132, 136, 138–140, 142, 143, 145, 147–149.

— потеря 472–473

— правдоподобия 241–242, 250

— — логарифмическая 252

— распределения вероятностей 90

— регрессии 352, 354

Х

Характеристики выборочные 355

— качества классификации 472

— — оценивания 234–240

— — статистического критерия 294–295

Характеристические (собственные) векторы 531, 532, 663

— — числа (значения) 530, 546, 663

Характеристические уравнения 530, 663

Ц

Целевая функция 368, 582

Целенаправленное проецирование 337, 339

Цензурирование выборки 262

Центральная предельная теорема 157, 166

— — — многомерная 158

Центральные моменты выборочные

207

— — случайной величины 100

Центрированное наблюдение 215, 541

Центроидный метод 561

Цепи Маркова 169

— — неприводимые 179

— — периодические 181

— — поглощающие 180

Цепочечный эффект 506

Ч

Частная (маргинальная) плотность вероятности 96

— — функция распределения 96

Частный коэффициент корреляции 418–420

— (маргинальный) закон распределения вероятностей 87

Частота (случаев) 202

— относительная 35, 37, 202

Число степеней свободы 142, 143, 144

Ш

Шкала измерений 362

— — количественная 362

— — номинальная (классификационная) 362

— — ординальная (порядковая) 362

Шкалирование многомерное — см.
Многомерное шкалирование

Э

Эвристические методы снижения размерности 570

Экзогенные переменные 350

Эконометрика 49

Эконометрическая модель 41

Экономико-математическая модель 41

Экспертно-статистический метод 368, 583

Экспертное упорядочение объектов 584

Экспоненциальное (показательное) распределение 137

Экстремальная группировка параметров 572

— постановка задач классификации

503

Экстремальные (оптимальные) свойства главных компонент 541

Эксцесс (распределения) 108, 210

Элементарное событие 53

Эмпирическая (выборочная) дисперсия 209

— — функция плотности 202

— — — распределения 201

Эмпирические аналоги начальных моментов 207

— — центральных моментов 207

— — центра группирования 208

Эндогенные переменные 350

Эргодическое свойство 181

Этапы вероятностно-статистического моделирования 43–45

— статистического исследования зависимостей 375–380

Эффективность оценки 238