математическая статистика

Математическая статистика

ИЗДАНИЕ ВТОРОЕ,
ПЕРЕРАБОТАННОЕ И ДОПОЛНЕННОЕ

го

МОСКВА «ВЫСШАЯ ШКОЛА» 1981

Стр.

ЛИТЕРАТУРА

1. *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. М., 1965.
2. *Венецкий И. Г.* Вариационные ряды и их характеристики. М., 1970.
3. *Венецкий И. Г., Кильдышев Г. С.* Теория вероятностей и математическая статистика. М., 1975.
4. *Вентцель Е. С.* Теория вероятностей. М., 1969.
5. *Вентцель Е. С., Овчаров Л. А.* Теория вероятностей. М., 1969.
6. *Гнеденко Б. В.* Курс теории вероятностей. М., 1965.
7. *Гнеденко Б. В.* Беседы о математической статистике. М., 1968.
8. *Гмурман В. Е.* Теория вероятностей и математическая статистика. М., 1972.
9. *Длин А. М.* Математическая статистика в технике. М., 1958.
10. *Дружинин Н. К.* Математическая статистика в экономике. М., 1971.
11. *Карасев А. И.* Теория вероятностей и математическая статистика. М., 1970.
12. *Митропольский А. К.* Техника статистических вычислений. М., 1971.
13. *Прохоров Ю. В., Розанов Ю. А.* Теория вероятностей. М., 1973.
14. *Пугачев В. С.* Введение в теорию вероятностей. М., 1968.
15. *Румшинский Л. З.* Элементы теории вероятностей. М., 1970.
16. *Свешников А. А.* Прикладные методы случайных функций. М., 1968.
17. *Смирнов Н. В., Дунин-Барковский И. В.* Курс теории вероятностей и математической статистики для технических приложений. М., 1965.
18. *Соболь И. М.* Метод Монте-Карло. М., 1978.
19. *Тутубалин В. Н.* Теория вероятностей. М., 1972.
20. *Фишер Р. А.* Статистические методы для исследователей. М., 1958.
21. *Хальд А.* Математическая статистика с техническими приложениями. М., 1956.
22. *Харман Г.* Современный факторный анализ. М., 1972.
23. *Шеффе Г.* Дисперсионный анализ. М., 1963.

Предисловие

Первое издание учебника для техникумов «Математическая статистика» под редакцией проф. А. М. Длина вышло в 1975 г.

В настоящее издание добавлены две главы: «Регрессионный анализ» и «Планирование эксперимента». Включен ряд новых тем: «Понятие о системе случайных величин», «Двумерное нормальное распределение», «Оценка значимости коэффициентов регрессии», «Факторный анализ» и т. д.; во многие параграфы внесены существенные изменения и уточнения.

В основу настоящего учебника, как и в основу учебника, вышедшего в 1975 г., положены методические принципы преподавания курса математической статистики, разработанные проф. А. М. Длином.

Введение, гл. 1, 12, 13 написаны В. Н. Калининой, гл. 2—4 — В. М. Ивановой, гл. 5—7 и 11 — Л. А. Нешумовой, гл. 8—10 — И. О. Решетниковой.

Авторы выражают благодарность Л. И. Трошину за рецензирование рукописи книги. Сделаные им замечания способствовали улучшению ее содержания. Авторы также будут признательны всем, кто пришлет свои критические замечания и пожелания, направленные на улучшение книги.

Авторы

Введение

Цель каждой науки — познание некоторых общих закономерностей, позволяющих предвидеть течение явлений и выбирать рациональные пути поведения в типичных ситуациях.

В основе научных знаний лежит наблюдение. Однако единичное наблюдение может нести много особенностей, свойственных только наблюдаемому единичному объекту и не отражать общей природы интересующего исследователя явления. Для обнаружения общей закономерности, которой подчиняется явление, необходимо многократно его наблюдать в одинаковых условиях. Чем это вызвано и что понимать под одинаковыми условиями?

Как известно, все явления окружающего мира взаимно связаны и влияют одно на другое. Каждое наблюдаемое явление связано причинной зависимостью с бесчисленным множеством других явлений, и течение его зависит от бесчисленного множества факторов. Проследить все это бесконечное множество связей и определить действие каждой из них невозможно. Поэтому ограничиваются изучением влияния лишь основных факторов, определяющих течение явления. При выяснении закономерностей этого влияния желательно, чтобы все прочие факторы от наблюдения к наблюдению сохранялись неизменными, но это также невозможно. Удается зафиксировать значения лишь некоторых контролируемых факторов. Под одинаковыми условиями наблюдений и понимается соблюдение во всех наблюдениях практически одинаковых значений контролируемых факторов.

Рассмотрим пример. Станок, хорошо отлаженный в начале работы, со временем теряет настройку, что приводит к ухудшению качества обработки изделий; к ухудшению качества приводит также и затупление режущего инструмента. Поставлена задача — угадать момент, когда следует остановить станок и провести его подналадку или сменить инструмент. Единственный способ определения этого момента — проверка качества изготовлен-

ных изделий. Наблюдение показало, что качество ухудшается после 2 ч работы режущего инструмента. Можно ли, основываясь на единственном наблюдении, планировать остановку станка по истечении 2 ч работы инструмента? Очевидно, нет. Для получения обоснованных выводов необходимы сведения по большему числу наблюдений, которые следует проводить в одинаковых условиях (должны быть одинаковыми качество режущих инструментов, заготовок и прочие контролируемые факторы). Как организовать сбор сведений? Сколько должно быть проведено наблюдений? Как обработать результаты наблюдений и сделать обоснованные практические выводы? Эти вопросы встают перед исследователем. Получить на них ответ позволяет специальный раздел математики — математическая статистика.

Рассмотрим еще один пример. Исследователя интересует зависимость урожайности от количества внесенных удобрений и качества обработки почвы. Для выяснения этой зависимости собраны сведения об урожайности, количестве внесенных удобрений и качестве обработки по достаточно большому числу одинаковых участков (с одинаковыми почвами, климатическими условиями, организацией работы по сбору урожая и т. д.). Как, используя эти сведения, количественно оценить зависимость урожайности от количества внесенных удобрений и качества обработки почвы и использовать ее для предвидения урожайности? На этот вопрос также дает ответ математическая статистика.

Итак, при изучении законов явлений можно зафиксировать на постоянном уровне лишь контролируемые факторы. Значения же неконтролируемых факторов от наблюдения к наблюдению меняются. Поэтому в любом реальном явлении всегда наблюдаются отклонения от существующих закономерностей. В этом смысле закономерность любого явления значительно «беднее», уже самого явления и не может характеризовать явление во всей полноте и многообразии. Наблюдаемые отклонения от закономерностей случайны: в одном наблюдении отклонение меньше, в другом — больше; предвидеть, каково отклонение в единичном наблюдении, нельзя.

Наблюдаемые в реальном явлении отклонения от закономерности, вызываемые совместным действием бесчисленного множества неучтенных факторов, представляют собой случайные явления. Какова роль случайных

явлений, случайных отклонений от закономерностей? Следует ли изучать их? В некоторых явлениях случайные отклонения от закономерностей настолько малы, что их можно не учитывать. Так обстоит дело со многими законами, известными из школьного курса физики и химии, например, законами Бойля — Мариотта, Гей-Люсака, Паскаля. При выполнении сформулированных в законах условий отклонения от этих законов практически отсутствуют. Есть и такие явления, в которых вообще невозможно подметить закономерностей, в которых случайность играет основную роль. Примером такого явления может служить движение малой частицы твердого вещества, взвешенной в жидкости (так называемое броуновское движение). Под действием толчков огромного количества движущихся молекул жидкости частица движется беспорядочно, без всякой видимой закономерности. В таких явлениях сама случайность является закономерностью.

При многократном наблюдении ряда случайных явлений в них самих можно заметить определенные закономерности. Изучив эти закономерности, человек получает возможность в определенном смысле управлять случайными явлениями, предвидеть их. Примером таких случайных явлений служат погрешности измерений. Измеряя один и тот же предмет, например взвешивая его на аналитических весах много раз, всегда получают близкие, но все же различные результаты. Это объясняется тем, что результат каждого измерения содержит случайную погрешность. Предвидеть эту погрешность, а следовательно, и результат каждого конкретного измерения нельзя. Однако если определенным образом систематизировать результаты измерений, то окажется, что в их изменении можно увидеть некоторую закономерность. Изучение этой закономерности по-прежнему не дает возможности предсказать результат единичного измерения, но позволяет, например, предвидеть в среднем результат серии измерений (а это уже немало!).

Из сказанного ясно, что изучению подлежат только такие случайные явления, которые можно, по крайней мере принципиально, наблюдать практически неограниченное число раз. Такие случайные явления называют *массовыми*. При этом наибольший интерес представляют массовые случайные явления, обладающие определенными закономерностями.

Математическая статистика — это раздел математики, изучающий методы сбора, систематизации и обработки результатов наблюдений массовых случайных явлений с целью выявления существующих закономерностей. Выводы о закономерностях, которым подчиняются явления, изучаемые методами математической статистики, всегда основываются на ограниченном, выборочном числе наблюдений. Поэтому естественно предположить, что эти выводы при большем числе наблюдений могут оказаться иными. Чтобы быть в состоянии высказать более определенное суждение об изучаемом явлении, математическая статистика опирается на теорию вероятностей.

В отличие от математической статистики, имеющей дело с результатами наблюдений над случайными явлениями, теория вероятностей формально-логически изучает закономерности случайных явлений и имеет дело с абстрактными описаниями действительности, с математическими моделями случайных явлений.

Оценив неизвестные величины или зависимости между ними по данным наблюдений, исследователь выдвигает ряд гипотез, предположений о том, что рассматриваемое явление можно описать той или иной вероятностной теоретической моделью. Далее, используя математико-статистические методы, можно дать ответ на вопрос, какую из гипотез или моделей следует принять. Именно эта модель и есть закономерность изучаемого явления. Таков типичный путь математико-статистического исследования.

Математическая статистика, опираясь при изучении закономерностей массовых, случайных явлений на вероятностные модели, в свою очередь, влияет на развитие теории вероятностей. Окружающий нас мир многообразен, и задачи, возникающие при изучении тех или иных случайных явлений, при обработке результатов наблюдений над ними требуют разработки новых вероятностных моделей. Математическая статистика и теория вероятностей — две неразрывно связанные науки.

Исследование задач, относящихся к массовым случайным явлениям, и появление соответствующего математического аппарата относятся к началу XVII в. Знаменитый физик Галилей исследует ошибки физических измерений, рассматривая их как случайные явления. Затем создается теория страхования, основанная на анали-

зе закономерностей таких массовых случайных явлений, как заболеваемость, смертность и т. д.

Возникновение теории вероятностей в современном смысле слова относится к середине XVII в. и связано с именами французских ученых Паскаля (1623—1662), Ферма (1601—1665) и Гюйгенса (1629—1695). Дальнейшим развитием теория вероятностей обязана Якову Бернулли (1654—1705), Муавру (1667—1754), Лапласу (1749—1827), Гауссу (1777—1855) и Пуассону (1781—1840).

Серьезные работы по теории ошибок измерений конца XVIII в., по обработке результатов биологических исследований в XIX в. позволили в начале XIX в. выделить математическую статистику в особую науку. В начальный период развития математической статистики большое значение имели работы А. Кетле (1796—1874), Ф. Гальтона (1822—1911) и в особенности К. Пирсона (1857—1936).

В 20—30-х годах XIX в. в России создается знаменитая Петербургская математическая школа, трудами которой теория вероятностей и математическая статистика были поставлены на прочную математическую основу и сделаны эффективными средствами познания. Среди ученых русской математической школы следует назвать в первую очередь П. Л. Чебышева (1821—1894), А. А. Маркова (1856—1922), А. М. Ляпунова (1857—1918).

Советская школа математической статистики и теории вероятностей занимает в мировой науке ведущее место. Основополагающие работы принадлежат А. Н. Колмогорову, А. Я. Хинчину (1894—1959), С. Н. Бернштейну, В. И. Романовскому (1879—1954), Н. В. Смирнову (1900—1966), Е. Е. Слуцкому (1880—1948), Б. В. Гнеденко, Ю. В. Линнику (1901—1972).

Развитие зарубежной теории вероятностей и математической статистики в настоящее время связано с именами английских ученых Стьюдента, Р. Фишера, Э. Пирсона, В. Феллера, Д. Дуба, Г. Крамера и американских Ю. Неймана и А. Вальда.

Глава 1

Построение вариационных рядов и вычисление статистических характеристик

§ 1.1. Вариационные ряды

Установление закономерностей, которым подчиняются массовые случайные явления, основано на изучении статистических данных — сведений о том, какие значения принял в результате наблюдений интересующий исследователя признак.

Пример 1.1. Исследователь, интересующийся тарифным разрядом рабочих механического цеха, в результате опроса 100 рабочих получил следующие сведения:

5, 1, 4, 5, 4, 3, 5, 5, 2, 5, 5, 6, 4, 3, 1, 5, 2, 5, 5, 5, 3, 3, 3, 6, 6, 5, 6, 5, 3, 4, 5, 4, 6, 6, 5, 2, 1, 5, 4, 5, 5, 3, 6, 4, 5, 5, 4, 3, 5, 5, 5, 4, 5, 6; 1, 5, 2, 6, 4, 4, 3, 5, 6, 3, 5, 6, 2, 5, 4, 5, 5, 4, 6, 5, 2, 5, 3, 4, 5, 6, 5, 5, 3, 5, 4, 6, 6, 5, 5, 4, 5, 5, 6, 5, 6, 5, 5, 6, 5, 5.

Здесь признаком является тарифный разряд, а полученные о нем сведения образуют статистические данные. Для изучения данных прежде всего необходимо их сгруппировать. Расположим наблюдавшиеся значения признака в порядке возрастания. Эта операция называется *ранжированием* статистических данных. В результате получим следующий ряд, который называется *ранжированным:*

$$\underbrace{1, 1, 1, 1}_{4\text{раза}}\ \underbrace{2, 2, 2, 2, 2, 2}_{6\text{раз}}\ \underbrace{3, 3, \ldots, 3}_{12\text{раз}}\ \underbrace{4, 4, \ldots, 4}_{16\text{раз}}$$

$$\underbrace{5, 5, \ldots, 5}_{44\text{раза}}\ \underbrace{6, 6, \ldots, 6}_{18\text{раз}}.$$

Из ранжированного ряда следует, что признак (тарифный разряд) принял шесть различных значений: первый, второй и т. д. до шестого разряда.

В дальнейшем различные значения признака условимся называть *вариантами,* а под варьированием — понимать изменение значений признака. Если признак по своей сущности таков, что различные его значения не могут отличаться друг от друга меньше чем на некоторую конечную величину, то говорят, что это *дискретно варьирующий признак.*

Тарифный разряд — дискретно варьирующий признак: его различные значения не могут отличаться друг от друга меньше, чем на единицу. В примере этот признак принял 6 различных значений — 6 вариантов: вариант 1 повторился 4 раза, вариант 2—6 раз и т. д. Число, показывающее, сколько раз встречается вариант x в ряде наблюдений, называется *частотой* варианта m_x. Ранжированный ряд представим в виде табл. 1.1.

Таблица 1.1

Тарифный разряд x	Количество рабочих m_x	Доля рабочих w_x
1	4	0,04
2	6	0,06
3	12	0,12
4	16	0,16
5	44	0,44
6	18	0,18
Σ	100	1,00

Вместо частоты варианта x можно рассматривать ее отношение к общему числу наблюдений n, которое называется *частостью* варианта x и обозначается w_x. Так как общее число наблюдений равно сумме частот всех вариантов ($n = \sum_x m_x$), то справедлива следующая цепочка равенств: $w_x = m_x/n = m_x/\Sigma m_x$.

Таблица, позволяющая судить о распределении частот (или частостей) между вариантами, называется *дискретным вариационным рядом.*

В примере 1.1 была поставлена задача изучить результаты наблюдений. Если просмотр первичных данных не позволил составить представление о варьировании

значений признака, то, рассматривая вариационный ряд, можно сделать следующие выводы: тарифный разряд колеблется от 1-го до 6-го; наиболее часто встречается 5-й тарифный разряд; с ростом тарифного разряда (до 5-го разряда) растет число рабочих, имеющих соответствующий разряд.

Наряду с понятием частоты используют понятие *накопленной частоты,* которую обозначают $m_x^{\text{нак}}$. Накопленная частота показывает, во скольких наблюдениях признак принял значения, меньшие заданного значения x. Отношение накопленной частоты к общему числу наблюдений называют *накопленной частостью* и обозначают $w_x^{\text{нк}}$. Очевидно, что $w_x^{\text{нак}} = m_x^{\text{нак}}/n$.

В дискретном вариационном ряду накопленные частоты (частости) вычисляются для каждого варианта и являются результатом последовательного суммирования частот (частостей). Накопленные частоты (частости) для вариационного ряда, заданного в табл. 1.1, вычислены в табл. 1.2.

Таблица 1.2

Тарифный разряд x	Количество рабочих m_x	Накопленная частота $m_x^{\text{нак}}$	Накопленная частость $w_x^{\text{нак}}$
1	4	0	0,00
2	6	0 + 4 = 4	0,04
3	12	4 + 6 = 10	0,10
4	16	10 + 12 = 22	0,22
5	44	22 + 16 = 38	0,38
6	18	38 + 44 = 82	0,82
Выше 6	0	82 + 18 = 100	1,00

Например, варианту 1 соответствует накопленная частота, равная нулю, так как среди опрошенных рабочих не было таких, у которых тарифный разряд был бы меньше 1-го; варианту 5 соответствует накопленная частота 38, так как было 4+6+12+16 рабочих с тарифным разрядом, меньшим 5-го, накопленная частость для этого варианта равна 0,38 (38 : 100); если тарифный разряд выше 6-го, то ему соответствует накопленная частота 100, так как тарифный разряд всех опрошенных рабочих не выше 6-го.

Пример 1.2. Исследователь, изучающий выработку на одного рабочего-станочника механического цеха в от-

четном году в процентах к предыдущему году, получил следующие данные (в целых процентах) по 117 рабочим:

111, 85, 85, 91, 101, 109, 86, 102, 111, 98, 105, 85, 112, 98, 112, 113, 87, 109, 109, 115, 99, 105, 111, 94, 107, 99, 107, 125, 89, 104, 113, 96, 103, 145, 104, 105, 88, 103, 97, 115, 109, 89, 108, 107, 97, 106, 107, 96, 109, 116, 109, 117, 108, 109, 139, 116, 117, 103, 127, 119, 118, 125, 105, 116, 117, 106, 101, 113, 107, 105, 119, 107,119, 111, 112, 129, 113, 106, 104, 106, 98, 123, 108, 93, 105, 106, 139, 108, 109, 93, 107, 117, 107, 118, 99, 108, 108, 119, 98, 108, 101, 109, 109, 128, 128, 127, 121, 118, 122, 116, 124, 125, 114, 126, 131, 141, 143.

В этом примере признаком является выработка в отчетном году в процентах к предыдущему. Очевидно, что значения, принимаемые этим признаком, могут отличаться одно от другого на сколь угодно малую величину, т. е. признак может принять любое значение в некотором числовом интервале (только для упрощения дальнейших расчетов полученные данные округлены до целых процентов). Такой признак называют *непрерывно варьирующим*. По приведенным данным трудно выявить характерные черты варьирования значений признака. Построение дискретного вариационного ряда также не даст желаемых результатов (слишком велико число различных наблюдавшихся значений признака). Для получения ясной картины объединим в группы рабочих, у которых величина выработки колеблется, например, в пределах 10%. Сгруппированные данные представим в табл. 1.3.

Таблица 1.3

Выработка в отчетном году в процентах к предыдущему интервалу	Количество рабочих m	Доля рабочих w	Накопленная частота $m_{нак}$	Накопленная частость $w_{нак}$
80—90	8	8/117	8	8/117
90—100	15	15/117	23	23/117
100—110	46	46/117	69	69/117
110—120	29	29/117	98	98/117
120—130	13	13/117	111	111/117
130—140	3	3/117	114	114/117
140—150	3	3/117	117	117/117
Σ	117	1		

В табл. 1.3 частоты m показывают, во скольких наблюдениях признак принял значения, принадлежащие тому или иному интервалу. Такую частоту называют *интервальной*, а отношение ее к общему числу наблюдений — *интервальной частостью* w. Таблицу, позволяющую судить о распределении частот (или частостей) между интервалами варьирования значений признака, называют *интервальным вариационным рядом*.

Интервальный вариационный ряд, представленный в табл. 1.3, позволяет выявить закономерности распределения рабочих по интервалам выработки. В табл. 1.3 для верхних границ интервалов приведены накопленные частоты (частости) (они получены последовательным суммированием интервальных частот (частостей), начиная с частоты (частости) первого интервала). Например, для верхней границы третьего интервала, равной 110, накопленная частота равна 69; так как 8+15+46 рабочих имели выработку меньше 110%, накопленная частость равна 69/117.

Интервальный вариационный ряд строят по данным наблюдений за непрерывно варьирующим признаком, а также за дискретно варьирующим, если велико число наблюдавшихся вариантов. Дискретный вариационный ряд строят только для дискретно варьирующего признака.

Иногда интервальный вариационный ряд условно заменяют дискретным. Тогда серединное значение интервала принимают за вариант x, а соответствующую интервальную частоту — за m_x.

§ 1.2. Построение интервального вариационного ряда

Для построения интервального вариационного ряда необходимо определить величину интервала, установить полную шкалу интервалов, в соответствии с ней сгруппировать результаты наблюдений. В примере 1.2 при выборе величины интервала учитывались требования наибольшего удобства отсчетов. Интервал был принят равным 10% и оказался удачным. Построенный интервальный ряд позволил выявить закономерности варьирования значений признака. Для определения оптимального интервала h, т. е. такого, при котором построенный интервальный ряд не был бы слишком громоздким и в

то же время позволял выявить характерные черты рассматриваемого явления, можно использовать *формулу Стэрджеса*

$$h = (x_{max} - x_{min})/(1 + 3{,}322 \lg n), \qquad (1.1)$$

где x_{max} и x_{min} — соответственно максимальный и минимальный варианты. Если h — дробное число, то за величину интервала следует взять либо ближайшее целое число, либо ближайшую несложную дробь.

За начало первого интервала рекомендуется принимать величину $a_1 = x_{min} - h/2$; начало второго интервала совпадает с концом первого и равно $a_2 = a_1 + h$; начало третьего интервала совпадает с концом второго и равно $a_3 = a_2 + h$. Построение интервалов продолжают до тех пор, пока начало следующего по порядку интервала не будет больше x_{max}.

После установления шкалы интервалов следует сгруппировать результаты наблюдений. Границы последовательных интервалов записывают в столбец слева, а затем, просматривая статистические данные в том порядке, в каком они были получены, проставляют черточки справа от соответствующего интервала. В интервал включаются данные, большие или равные нижней границе интервала и меньшие верхней границы. Целесообразно каждые пятое и шестое наблюдения отмечать диагональными черточками, пересекающими квадрат из четырех предшествующих. Общее количество черточек, проставленных против какого-либо интервала, определяет его частоту.

Пример 1.3. Построить интервальный вариационный ряд по следующим результатам измерения диаметра валика после шлифовки:

6,75; 6,77; 6,77; 6,73; 6,76; 6,74; 6,70; 6,75; 6,71; 6,72; 6,77; 6,79; 6,71; 6,78; 6,73; 6,70; 6,73; 6,77; 6,75; 6,74; 6,71; 6,70; 6,78; 6,76; 6,81; 6,69; 6,80; 6,80; 6,77; 6,68; 6,74; 6,70; 6,70; 6,74; 6,77; 6,83; 6,76; 6,76; 6,82; 6,77; 6,71; 6,74; 6,77; 6,75; 6,74; 6,75; 6,77; 6,72; 6,74; 6,80; 6,75; 6,80; 6,72; 6,78; 6,70; 6,75; 6,78; 6,78; 6,76; 6,77; 6,74; 6,74; 6,77; 6,73; 6,74; 6,77; 6,74; 6,75; 6,74; 6,76; 6,76; 6,74; 6,74; 6,74; 6,74; 6,76; 6,74; 6,72; 6,80; 6,76; 6,78; 6,73; 6,70; 6,76; 6,76; 6,77; 6,75; 6,78; 6,72; 6,76; 6,78; 6,68; 6,75; 6,73; 6,82; 6,73; 6,80; 6,81; 6,71; 6,82; 6,77; 6,80; 6,80; 6,70; 6,70; 6,82; 6,72; 6,69; 6,73; 6,76; 6,74; 6,77; 6,72; 6,76; 6,78; 6,78; 6,73; 6,76; 6,80; 6,76; 6,72; 6,76; 6,76; 6,70; 6,73; 6,75; 6,77; 6,77; 6,70; 6,81; 6,74; 6,73; 6,77; 6,74; 6,78; 6,69; 6,74; 6,71. 6,76; 6,76; 6,77; 6,70; 6,81; 6,74; 6,74; 6,77; 6,75; 6,80; 6,74; 6,76; 6,77; о,77; 6,81; 6,75; 6,78; 6,73; 6,76; 6,76; 6,76; 6,77; 6,76; 6,80; 6,77; 6,74; 6,77; 6,72; 6,75; 6,76; 6,77; 6,81; 6,76; 6,76; 8,76; 6,80; 6,74; 6,80; 6,74; 6,73; 6,75; 6,77; 6,74; 6,76; 6,77; 6,77; 6,75; 6,76; 6,74; 6,82; 6,76; 6,73; 6,74; 6,75; 6,76; 6,72; 6,78; 6,72; 6,76; 6,77; 6,75; 6,78.

Решение. Величину интервала определим по формуле (1.1): $h = (6{,}83 - 6{,}68)/(1 + 3{,}322 \lg 200) \approx 0{,}15/8{,}64 \approx 0{,}02$. За начало первого интервала принимаем величину $a_1 = 6{,}68 - 0{,}01 = 6{,}67$. Тогда $a_2 = 6{,}67 + 0{,}02 = 6{,}69$; $a_3 = 6{,}71$ и т. д. Шкала интервалов и группировка результатов наблюдений приведены в табл. 1.4.

Таблица 1.4

Диаметр валика после шлифовки (интервалы), мм	Подсчет частот	Частота m	Накопленная частота m
6,67—6,69		2	2
6,69—6,71		15	17
6,71—6,73		17	34
6,73—6,75		44	78
6,75—6,77		52	130
6,77—6,79		44	174
6,79—6,81		14	188
6,81—6,83		11	199
6,83—6,85		1	200
Σ		200	

§ 1.3. Графическое изображение вариационных рядов

Графическое изображение вариационного ряда позволяет представить в наглядной форме закономерности варьирования значений признака. Наиболее широко используются следующие виды графического изображения вариационных рядов: полигон, гистограмма, кумулятивная кривая.

Полигон, как правило, служит для изображения дискретного вариационного ряда. Для его построения в прямоугольной системе координат наносят точки с координатами $(x;\ m_x)$, где x — варианта, а m_x — соответствующая ему частота. Иногда вместо точек $(x;\ m_x)$ строят точки $(x;\ w_x)$. Затем эти точки соединяют последовательно отрезками. Крайние левую и правую точки соединяют соответственно с точками, изображающими ближайший снизу к наименьшему и ближайший сверху к наибольшему варианты. Полученная ломаная линия называется полигоном.

Пример 1.4. Построить полигон распределения по данным, приведенным в табл. 1.1.

Решение. См. рис. 1.1.

Гистограмма служит для изображения только интервального вариационного ряда. Для ее построения в прямоугольной системе координат по оси абсцисс откладывают отрезки, изображающие интервалы варьирования, и на этих отрезках, как на основании, строят прямоугольники с высотами, равными частотам (или частостям) соответствующего интервала. В результате получают ступенчатую фигуру, состоящую из прямоугольников, которую и называют *гистограммой*.

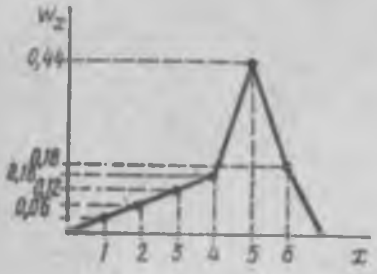

Рис. 1.1

Если по оси абсцисс выбрать такой масштаб, чтобы ширина интервала была равна единице, и считать, что по оси ординат единица масштаба соответствует одному наблюдению, то площадь гистограммы равна общему числу наблюдений, если по оси ординат откладывались частоты, и эта площадь равна единице, если откладывались частости.

Пример 1.5. Построить гистограмму распределения по данным, приведенным в табл. 1.3.

Решение. См. рис. 1.2.

Иногда интервальный ряд изображают с помощью полигона. В этом случае интервалы заменяют их серединными значениями и к ним относят интервальные частоты. Для полученного дискретного ряда строят полигон.

Пример 1.6. Построить полигон распределения по данным, приведенным в табл. 1.2.

Решение. См. рис. 1.2.

Кумулятивная кривая (кривая накопленных частот или накопленных частостей) строится следующим образом. Если вариационный ряд дискретный, то в прямоугольной системе координат строят точки с координатами $(x; m_x^{\text{нак}})$, где x — варианта, $m_x^{\text{нак}}$ — соответствующая накопленная частота. Иногда вместо точек $(x; m_x^{\text{нак}})$ строят точки $(x; w_x^{\text{нак}})$. Полученные точки соединяют отрезками.

Если вариационный ряд интервальный, то по оси абсцисс откладывают интервалы. Верхним границам интервалов соответствуют накопленные частоты (или накопленные частости); нижней границе первого интервала — накопленная частота, равная нулю. Построив кумулятивную кривую, можно приблизительно установить число наблюдений (или их долю в общем количестве наблю-

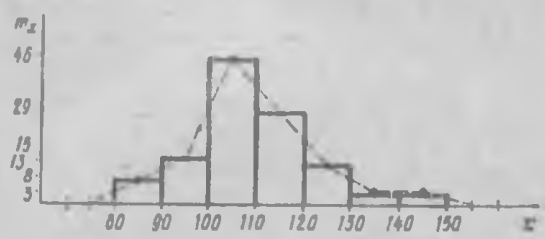

Рис. 1.2.

дений), в которых признак принял значения, меньшие заданного.

Пример 1.7. Построить кумулятивную кривую по данным, приведенным в табл. 1.2.

Решение. См. рис. 1.3.

Пример 1.8. Построить кумулятивную кривую по данным, приведенным в табл. 1.3.

Решение. См. рис. 1.4.

Построение вариационного ряда — первый шаг к осмыслению ряда наблюдений. Однако на практике это-

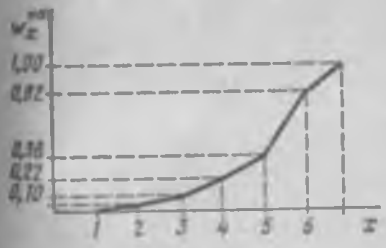

Рис. 1.3

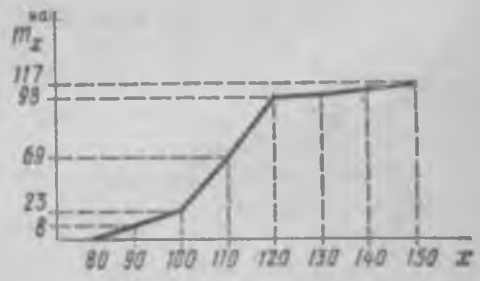

Рис. 1.4

го недостаточно, особенно когда необходимо сравнить два ряда или более. Сравнению подлежат только так называемые *однотипные* вариационные ряды, т. е. ряды, которые построены по результатам обработки сходных статистических данных. Например, можно сравнивать

распределения рабочих по возрасту на двух заводах или распределения времени простоев станков одного вида. Однотипные вариационные ряды обычно имеют похожую форму при графическом изображении, однако могут отличаться друг от друга, а именно: иметь различные значения признака, вокруг которых концентрируются наблюдения (меры этой качественной особенности назвася *средними величинами);* различаться рассеянием наблюдений вокруг средних величин (меры этой особенности получили название *показателей вариации).*

Средние величины и показатели вариации позволяют судить о характерных особенностях вариационного ряда и называются *статистическими характеристиками.* К статистическим характеристикам относятся также показатели, характеризующие различия в скошенности полигонов и различия в их островершинности.

§ 1.4. Средние величины

Средние величины являются как бы «представителями» всего ряда наблюдений, поскольку вокруг них концентрируются наблюдавшиеся значения признака. Заметим, что только для качественно однородных наблюдений имеет смысл вычислять средние величины.

Различают несколько видов средних величин: средняя арифметическая, средняя геометрическая, средняя гармоническая, средняя квадратическая, средняя кубическая и т. д. При выборе вида средней величины необходимо прежде всего ответить на вопрос: какое свойство ряда мы хотим представить средней величиной или, иначе говоря, какая цель преследуется при вычислении средней? Это свойство, получившее название *определяющего,* и определяет вид средней. Понятие определяющего свойства впервые введено советским статистиком А. Я. Боярским.

Наиболее распространенной средней величиной является средняя арифметическая. Пусть $x_1, x_2, \ldots, x_n$ — данные наблюдений; $\bar{x}$ — средняя арифметическая. Свойство, определяющее среднюю арифметическую, формулируется следующим образом: *сумма результатов наблюдений должна остаться неизменной, если каждое из них заменить средней арифметической, т. е.*

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \bar{x}. \qquad (1.2)$$

Так как $\bar{x}=\text{const}$, то $\sum_{i=1}^{n} x_i = n\bar{x}$. Отсюда получаем следующую формулу для вычисления средней арифметической по данным наблюдений:

$$\bar{x} = \left(\sum_{i=1}^{n} x_i\right) \Big/ n. \qquad (1.3)$$

Если по наблюдениям построен вариационный ряд, то средняя арифметическая

$$\bar{x} = \sum_{x} x m_x \Big/ \sum_{x} m_x, \qquad (1.4)$$

где x — вариант, если ряд дискретный, и центр интервала, если ряд интервальный; m_x — соответствующая частота.

Частоты m_x в формуле (1.4) называют *весами,* а операцию умножения x на m_x — *операцией взвешивания.* Среднюю арифметическую, вычисленную по формуле (1.4), называют *взвешенной* в отличие от средней арифметической, вычисленной по формуле (1.3).

Очевидно, что если по данным наблюдений построен дискретный вариационный ряд, то формулы (1.3) и (1.4) дают одинаковые значения средней арифметической. Если же по наблюдениям построен интервальный ряд, то средние арифметические, вычисленные по формулам (1.3) и (1.4), могут не совпадать, так как в формуле (1.4) значения признака внутри каждого интервала принимаются равными центрам интервалов. Ошибка, возникающая в результате такой замены, вообще говоря, очень мала, если наблюдения распределены равномерно вдоль каждого интервала, а не скапливаются к одноименным границам интервалов (т. е. либо все к нижним границам, либо все к верхним границам).

Среднюю арифметическую для вариационного ряда можно вычислять по формуле

$$\bar{x} = \sum_{x} x w_x, \qquad (1.5)$$

которая является следствием формулы (1.4). Действительно,

$$\bar{x} = \sum_{x} x m_x \Big/ \sum_{x} m_x = \sum_{x} x m_x \Big/ \sum_{n} m_x = \sum_{n} x w_x.$$

Пример 1.9. По данным, приведенным в примере 1.2, вычислить среднюю выработку одного рабочего в отчетном году в процентах к предыдущему.

Р е ш е н и е. Очевидно, что средняя выработка должна быть такой, чтобы сумма выработок всех 117 рабочих осталась неизменной, если выработку каждого из них заменить средней. Следовательно, средняя выработка есть средняя арифметическая. Воспользуемся сначала формулой (1.3). Сложим выработки всех рабочих и полученную сумму разделим на количество рабочих:

$$\bar{x} = (111+85+85+91+101+\cdots+141+143)/117 \approx 109{,}7\%.$$

Теперь вычислим взвешенную среднюю арифметическую по формуле (1.4), используя вариационный ряд, заданный табл. 1.3:

$$\bar{x} = \frac{85\cdot 8+95\cdot 15+105\cdot 46+115\cdot 29+125\cdot 13+135\cdot 3+145\cdot 3}{8+15+46+29+13+3+3} \approx 108{,}8\%.$$

Ошибка взвешенной средней арифметической, возникшая вследствие замены результатов наблюдений серединами соответствующих интервалов, составляет 0,9%.

Заметим, что при вычислении средней выработки операция взвешивания заключается в умножении выработки на число рабочих, в результате получаем общую выработку рабочих. В дальнейшем, какая бы средняя ни вычислялась, необходимо помнить, что *операция взвешивания должна иметь конкретный смысл.*

Пример 1.10. Вычислить средний тарифный разряд рабочих-станочников по вариационному ряду, заданному табл. 1.2. При решении воспользоваться формулой (1.5).

Р е ш е н и е. Средний тарифный разряд

$$\bar{x} = 1\cdot 0{,}04 + 2\cdot 0{,}06 + 3\cdot 0{,}12 + 4\cdot 0{,}16 + 5\cdot 0{,}44 + 6\cdot 0{,}18 \approx 4.$$

Свойство, определяющее среднюю арифметическую, сводилось к требованию неизменности суммы наблюдений при замене каждого из них средней арифметической. При решении практических задач может оказаться необходимым вычислить такую среднюю $\bar{x}_q$, при замене которой каждого наблюдения, осталась бы неизменной сумма q-х степеней наблюдений, т. е. чтобы

$$\sum_{i=1}^{n} x_i^q = \sum_{i=1}^{n} (\bar{x}_q)^q, \tag{1.6}$$

где q — положительное или отрицательное число. Среднюю $\bar{x}_q$ называют *степенной средней q-го порядка.* Из определяющего свойства (1.6) получим следующую формулу для вычисления $\bar{x}_q$ по данным наблюдений:

$$\bar{x}_q = \sqrt[q]{\sum_{i=1}^{n} x_i^q / n}. \tag{1.7}$$

Сравнивая формулы (1.7) и (1.3), можно сделать вывод, что степенная средняя первого порядка есть не что иное, как средняя арифметическая, т. е. $\bar{x}_1 = \bar{x}$.

При $q=-1$ из формулы (1.7) получаем выражение для средней гармонической, при $q=2$ — для средней квадратической, при $q=3$ — для средней кубической и т. д.

Средней геометрической $\bar{x}_{\text{геом}}$ называют корень n-й степени из произведения наблюдений $x_1, x_2, \ldots, x_n$: $\bar{x}_{\text{геом}} = \sqrt[n]{\prod_{i=1}^{n} x_i}$. Можно доказать, что средняя геометрическая является предельным случаем степенной средней q-го порядка при $q=0$, т. е.

$$\bar{x}_{\text{геом}} = \lim_{q \to 0} \bar{x}_q.$$

Рассмотрим основные свойства средней арифметической.

1°. *Сумма отклонений результатов наблюдений от средней арифметической равна нулю.*

Доказательство. Исходя из определяющего свойства (1.2) средней арифметической, получаем

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

Если по результатам наблюдений построен вариационный ряд и средняя арифметическая взвешенная, то свойство 1° формулируется так: *сумма произведений отклонений вариантов от средней арифметической на соответствующие частоты равна нулю.* Действительно, на основании формулы (1.4) получаем

$$\sum_x x m_x = \bar{x} \sum_x m_x = \sum_x \bar{x} m_x,$$

или

$$0 = \sum_x x m_x - \sum_x \bar{x} m_x = \sum_x (x - \bar{x}) m_x. \qquad \square$$

2°. *Если все результаты наблюдений уменьшить (увеличить) на одно и то же число, то средняя арифметическая уменьшится (увеличится) на то же число**.

Доказательство. Очевидно, что при уменьшении вариантов на одно и то же число c соответствующие им частоты останутся прежними. Поэтому взвешенная средняя арифметическая для измененного вариационного ряда такова:

$$\overline{x - c} = \sum_x (x - c) m_x \Big/ \sum_x m_x =$$

$$= \left(\sum_x x m_x - \sum_x c m_x \right) \Big/ \sum_x m_x =$$

* Доказательство свойств 2° и 3° проведем в предположении, что по результатам наблюдений построен вариационный ряд и средняя арифметическая — взвешенная.

$$= \sum_x xm_x \Big/ \left(\sum_x m_x - \sum_x cm_x \right) \Big/ \sum_x m_x = \bar{x} - c.$$

Аналогично можно показать, что $\overline{x+c} = \bar{x} + c$. Это свойство позволяет среднюю арифметическую вычислять не по данным вариантам, а по уменьшенным (увеличенным) на одно и то же число c. Если среднюю арифметическую, вычисленную для измененного ряда, увеличить (уменьшить) на число c, то получим среднюю арифметическую для первоначального вариационного ряда.

3°. *Если все результаты наблюдений уменьшить (увеличить) в одно и то же число раз, то средняя арифметическая уменьшится (увеличится) во столько же раз.*

Доказательство. Очевидно, что при уменьшении вариантов в k раз их частоты останутся прежними. Поэтому средняя арифметическая для измененного ряда

$$\overline{\left(\frac{x}{k}\right)} = \frac{\sum_x (x/k)\, m_x}{\sum_x m_x} = \frac{1}{k} \frac{\sum_x xm_x}{\sum_x m_x} = \frac{\bar{x}}{k}. \quad \square$$

Аналогично можно доказать, что $\overline{(xk)} = \bar{x} \cdot k$. Рассмотренное свойство позволяет среднюю арифметическую вычислять не по данным вариантам, а по уменьшенным (увеличенным) в одно и то же число k раз. Если среднюю арифметическую, вычисленную для измененного ряда, увеличить (уменьшить) в k раз, то получим среднюю арифметическую для первоначального вариационного ряда.

4°. *Если ряд наблюдений состоит из двух групп наблюдений, то средняя арифметическая всего ряда равна взвешенной средней арифметической групповых средних, причем весами являются объемы групп.*

Пусть n_1 и n_2 — число наблюдений соответственно в 1-й и 2-й группах; $\bar{x}$ — средняя арифметическая для всего ряда (n_1+n_2) наблюдений; $\bar{x}_1$ и $\bar{x}_2$ — средние арифметические соответственно для 1-й и 2-й групп наблюдений. Требуется доказать, что

$$\bar{x} = (\bar{x}_1 n_1 + \bar{x}_2 n_2)/(n_1 + n_2).$$

Доказательство. Исходя из определяющего свойства средней арифметической, имеем: произведение $(n_1+n_2)\bar{x}$ равно сумме (n_1+n_2) наблюдавшихся значе-

ний признака; $n_1\bar{x}_1$ равно сумме n_1 наблюдавшихся значений, образующих первую группу: $n_2\bar{x}_2$ равно сумме n_2 наблюдавшихся значений, образующих вторую группу. Следовательно,

$$(n_1 + n_2)\bar{x} = n_1\bar{x}_1 + n_2\bar{x}_2. \quad \square$$

Следствие. *Если ряд наблюдений состоит из k групп наблюдений, то средняя арифметическая всего ряда $(\bar{x})$ равна взвешенной средней арифметической групповых средних $(\bar{x}_i)$, причем весами являются объемы групп (n_i):*

$$\bar{x} = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2 + \ldots + \bar{x}_k n_k}{n_1 + n_2 + \ldots + n_k} = \sum_{i=1}^{k} \bar{x}_i n_i \Big/ \sum_{i=1}^{k} n_i.$$

5°. *Средняя арифметическая для сумм (разностей) взаимно соответствующих значений признака двух рядов наблюдений с одинаковым числом наблюдений равна сумме (разности) средних арифметических этих рядов.*

Пусть $x_1, x_2, \ldots, x_n$ — один ряд наблюдений, $\bar{x}$ — его средняя арифметическая; $y_1, y_2, \ldots, y_n$ — другой ряд наблюдений, $\bar{y}$ — его средняя арифметическая; $x_1+y_1, x_2+y_2, \ldots, x_n+y_n$ — ряд сумм соответствующих наблюдений, $\overline{x+y}$ — его средняя арифметическая. Требуется доказать, что $\overline{x+y}=\bar{x}+\bar{y}$.

Доказательство. Имеем

$$\overline{x+y} = \frac{\sum_{i=1}^{n}(x_i + y_i)}{n} = \frac{\sum_{i=1}^{n} x_i}{n} + \frac{\sum_{i=1}^{n} y_i}{n} = \bar{x} + \bar{y}. \quad \square$$

Аналогично можно показать, что $\overline{x-y}=\bar{x}-\bar{y}$.

Следствие. *Средняя арифметическая алгебраической суммы соответствующих значений признака нескольких рядов наблюдений с одинаковым числом наблюдений равна алгебраической сумме средних арифметических этих рядов.*

Вычисление средней арифметической вариационного ряда непосредственно по формуле (1.4) приводит к громоздким расчетам, если числовые значения вариантов и соответствующие им частоты велики. Поэтому часто используют следующий способ, основанный на свойствах 3° и 2° средней арифметической: среднюю вычисляют не по первоначальным вариантам x_i, а по уменьшенным на не-

$$= \sum_x xm_x \Big/ \Big(\sum_x m_x - \sum_x cm_x \Big) \Big/ \sum_x m_x = \bar{x} - c.$$

Аналогично можно показать, что $\overline{x+c} = \bar{x} + c$. Это свойство позволяет среднюю арифметическую вычислять не по данным вариантам, а по уменьшенным (увеличенным) на одно и то же число c. Если среднюю арифметическую, вычисленную для измененного ряда, увеличить (уменьшить) на число c, то получим среднюю арифметическую для первоначального вариационного ряда.

3°. *Если все результаты наблюдений уменьшить (увеличить) в одно и то же число раз, то средняя арифметическая уменьшится (увеличится) во столько же раз.*

Доказательство. Очевидно, что при уменьшении вариантов в k раз их частоты останутся прежними. Поэтому средняя арифметическая для измененного ряда

$$\overline{\left(\frac{x}{k}\right)} = \frac{\sum_x (x/k)\, m_x}{\sum_x m_x} = \frac{1}{k} \frac{\sum_x xm_x}{\sum_x m_x} = \frac{\bar{x}}{k}. \quad \square$$

Аналогично можно доказать, что $\overline{(xk)} = \bar{x}\, k$. Рассмотренное свойство позволяет среднюю арифметическую вычислять не по данным вариантам, а по уменьшенным (увеличенным) в одно и то же число k раз. Если среднюю арифметическую, вычисленную для измененного ряда, увеличить (уменьшить) в k раз, то получим среднюю арифметическую для первоначального вариационного ряда.

4°. *Если ряд наблюдений состоит из двух групп наблюдений, то средняя арифметическая всего ряда равна взвешенной средней арифметической групповых средних, причем весами являются объемы групп.*

Пусть n_1 и n_2 — число наблюдений соответственно в 1-й и 2-й группах; $\bar{x}$ — средняя арифметическая для всего ряда (n_1+n_2) наблюдений; $\bar{x}_1$ и $\bar{x}_2$ — средние арифметические соответственно для 1-й и 2-й групп наблюдений. Требуется доказать, что

$$\bar{x} = (\bar{x}_1 n_1 + \bar{x}_2 n_2)/(n_1 + n_2).$$

Доказательство. Исходя из определяющего свойства средней арифметической, имеем: произведение $(n_1+n_2)\bar{x}$ равно сумме (n_1+n_2) наблюдавшихся значе-

ний признака; $n_1\bar{x}_1$ равно сумме n_1 наблюдавшихся значений, образующих первую группу: $n_2\bar{x}_2$ равно сумме n_2 наблюдавшихся значений, образующих вторую группу. Следовательно,

$$(n_1 + n_2)\bar{x} = n_1\bar{x}_1 + n_2\bar{x}_2. \quad \square$$

Следствие. *Если ряд наблюдений состоит из k групп наблюдений, то средняя арифметическая всего ряда $(\bar{x})$ равна взвешенной средней арифметической групповых средних $(\bar{x}_i)$, причем весами являются объемы групп (n_i):*

$$\bar{x} = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2 + \ldots + \bar{x}_k n_k}{n_1 + n_2 + \ldots + n_k} = \sum_{i=1}^{k} \bar{x}_i n_i \Big/ \sum_{i=1}^{k} n_i.$$

5°. *Средняя арифметическая для сумм (разностей) взаимно соответствующих значений признака двух рядов наблюдений с одинаковым числом наблюдений равна сумме (разности) средних арифметических этих рядов.*

Пусть $x_1, x_2, \ldots, x_n$ — один ряд наблюдений, $\bar{x}$ — его средняя арифметическая; $y_1, y_2, \ldots, y_n$ — другой ряд наблюдений, $\bar{y}$ — его средняя арифметическая; $x_1+y_1, x_2+y_2, \ldots, x_n+y_n$ — ряд сумм соответствующих наблюдений, $\overline{x+y}$ — его средняя арифметическая. Требуется доказать, что $\overline{x+y}=\bar{x}+\bar{y}$.

Доказательство. Имеем

$$\overline{x+y} = \frac{\sum_{i=1}^{n}(x_i + y_i)}{n} = \frac{\sum_{i=1}^{n} x_i}{n} + \frac{\sum_{i=1}^{n} y_i}{n} = \bar{x} + \bar{y}. \quad \square$$

Аналогично можно показать, что $\overline{x-y}=\bar{x}-\bar{y}$.

Следствие. *Средняя арифметическая алгебраической суммы соответствующих значений признака нескольких рядов наблюдений с одинаковым числом наблюдений равна алгебраической сумме средних арифметических этих рядов.*

Вычисление средней арифметической вариационного ряда непосредственно по формуле (1.4) приводит к громоздким расчетам, если числовые значения вариантов и соответствующие им частоты велики. Поэтому часто используют следующий способ, основанный на свойствах 3° и 2° средней арифметической: среднюю вычисляют не по первоначальным вариантам x, а по уменьшенным на не-

которое число c, а затем разделенным на некоторое число k, т. е. для вариантов $x' = (x - c)/k$. Зная среднюю арифметическую $\overline{x'}$ для измененного ряда, легко вычислить среднюю арифметическую для первоначального ряда:

$$\overline{x} = \overline{x'} \cdot k + c. \qquad (1.8)$$

Действительно, принимая во внимание свойства 3° и 2° средней арифметической, получаем

$$\overline{x'} = (\overline{(x - c)}/k) = (\overline{x - c})/k = (\overline{x} - c)/k,$$

откуда следует, что $\overline{x} = \overline{x'} \cdot k + c$.

Очевидно, что от выбора числовых значений c и k зависит, насколько простым будет вычисление средней арифметической для измененного ряда. Значения c и k обычно выбирают так, чтобы новые варианты $x' = (x - c)/k$ были небольшими целыми числами. Если ряд дискретный, то в качестве c берется вариант, занимающий серединное положение в вариационном ряду (если таких вариантов два, то за c принимается тот, которому соответствует большая частота); за k принимают наибольший общий делитель вариантов $(x - c)$. Если ряд интервальный, то его заменяют дискретным: тогда c — центр серединного интервала (если таких интервала два, то берется тот, которому соответствует большая частота); за k принимают длину интервала h.

Пример 1.11. Используя предложенный способ вычисления средней арифметической, по данным, приведенным в табл. 1.3, найти среднюю выработку одного рабочего в отчетном году в процентах к предыдущему.

Решение. Примем $c = 115$, $k = 10$. Результаты вычислений расположим в табл. 1.5.

Таблица 1.5

Выработка в отчетном году в процентах к предыдущему (интервалы)	Количество рабочих m_x	Середина интервала x	$x' = \frac{x-c}{k}$	$x'm_x$	$x'+1$	$(x'+1)\,m_x$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
80—90	8	85	—3	—24	—2	—16
90—100	15	95	—2	—30	—1	—15
100—110	46	105	—1	—46	0	0
110—120	29	115	0	0	1	29
120—130	13	125	1	13	2	26
130—140	3	135	2	6	3	9
140—150	3	145	3	9	4	12
Σ	117	—	—	—72	—	45

В соответствии с формулой (1.4) имеем

$$\overline{x'} = \sum x' m_x / \sum m_x = -72/117 \approx -0{,}62.$$

По формуле (1.8) находим $\bar{x} = -0{,}62\ 10 + 115 = 108{,}8\%$.

Для проверки хода вычислений заполняем два последних столбца таблицы. Так как $\Sigma(x'+1)m_x = \Sigma x' m_x + \Sigma m_x$, то $\Sigma_7 = \Sigma_5 + \Sigma_2$. В самом деле, $45 = (-72) + 117$.

§ 1.5. Медиана и мода

Наряду со средними величинами в качестве описательных характеристик вариационного ряда применяют медиану и моду.

Медианой ($\tilde{M}_e$) называют значение признака, приходящееся на середину ранжированного ряда наблюдений.

Пусть проведено нечетное число наблюдений, т. е. $n = 2q - 1$, и результаты наблюдений проранжированы и выписаны в следующий ряд: $x_{(1)}, x_{(2)}, \ldots, x_{(q-1)}, x_{(q)}, x_{(q+1)}, \ldots, x_{(n)}$. Здесь $x_{(i)}$ — значение признака, занявшее i-е порядковое место в ранжированном ряду. На середину ряда приходится значение $x_{(q)}$. Следовательно,

$$\tilde{M}_e = x_{(q)}$$

Если проведено четное число наблюдений, т. е. $n = 2q$, то на середину ранжированного ряда $x_{(1)}, \ldots, x_{(q)}, x_{(q+1)}, \ldots, x_{(n)}$ приходятся значения $x_{(q)}$ и $x_{(q+1)}$. В этом случае за медиану принимают среднюю арифметическую значений $x_{(q)}$ и $x_{(q+1)}$, т. е.

$$\tilde{M}_e = (x_{(q)} + x_{(q+1)})/2.$$

Покажем на примерах, как определяется медиана дискретного и интервального вариационных рядов.

Пример 1.12. Вычислить медиану дискретного вариационного ряда, заданного табл. 1.1.

Р е ш е н и е. Проведено четное число наблюдений: $n = 100 = 2 \cdot 50$, следовательно,

$$\tilde{M}_e = (x_{(50)} + x_{(51)})/2,$$

где $x_{(50)}$ и $x_{(51)}$ — значения, расположенные соответственно на 50-м и 51-м порядковых местах в ранжированном ряду наблюдений. Чтобы найти $x_{(50)}$ и $x_{(51)}$ по вариационному ряду, воспользуемся тем фактом, что накопленная частота — это число, показывающее, сколько наблюдалось значений признака, меньших данного. По ряду накопленных частот (см. табл. 1.2) видим, что наблюдалось 38 значений признака, меньших пяти, и 82 значения, меньших шести. Следовательно, $x_{(50)} = x_{(51)} = 5$, а медиана $\tilde{M}_e = (5+5)/2 = 5$ (тарифный разряд).

Пример 1.13. Вычислить медиану интервального вариационного ряда, заданного табл. 1.3.

Р е ш е н и е. Проведено нечетное число наблюдений $n = 117$. Так

которое число c, а затем разделенным на некоторое число k, т. е. для вариантов $x' = (x - c)/k$. Зная среднюю арифметическую $\overline{x'}$ для измененного ряда, легко вычислить среднюю арифметическую для первоначального ряда:

$$\overline{x} = \overline{x'} \cdot k + c. \qquad (1.8)$$

Действительно, принимая во внимание свойства 3° и 2° средней арифметической, получаем

$$\overline{x'} = (\overline{(x - c)}/k) = (\overline{x - c})/k = (\overline{x} - c)/k,$$

откуда следует, что $\overline{x} = \overline{x'} \cdot k + c$.

Очевидно, что от выбора числовых значений c и k зависит, насколько простым будет вычисление средней арифметической для измененного ряда. Значения c и k обычно выбирают так, чтобы новые варианты $x' = (x - c)/k$ были небольшими целыми числами. Если ряд дискретный, то в качестве c берется вариант, занимающий серединное положение в вариационном ряду (если таких вариантов два, то за c принимается тот, которому соответствует большая частота); за k принимают наибольший общий делитель вариантов $(x - c)$. Если ряд интервальный, то его заменяют дискретным: тогда c — центр серединного интервала (если таких интервала два, то берется тот, которому соответствует большая частота); за k принимают длину интервала h.

Пример 1.11. Используя предложенный способ вычисления средней арифметической, по данным, приведенным в табл. 1.3, найти среднюю выработку одного рабочего в отчетном году в процентах к предыдущему.

Решение. Примем $c = 115$, $k = 10$. Результаты вычислений расположим в табл. 1.5.

Таблица 1.5

Выработка в отчетном году в процентах к предыдущему (интервалы)	Количество рабочих m_x	Середина интервала x	$x' = \frac{x-c}{k}$	$x' m_x$	$x'+1$	$(x'+1)\, m_x$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
80—90	8	85	—3	—24	—2	—16
90—100	15	95	—2	—30	—1	—15
100—110	46	105	—1	—46	0	0
110—120	29	115	0	0	1	29
120—130	13	125	1	13	2	26
130—140	3	135	2	6	3	9
140—150	3	145	3	9	4	12
Σ	117	—	—	—72	—	45

В соответствии с формулой (1.4) имеем

$$\overline{x'} = \sum x' m_x / \sum m_x = -72/117 \approx -0{,}62.$$

По формуле (1.8) находим $\bar{x} = -0{,}62 \cdot 10 + 115 = 108{,}8\%$.

Для проверки хода вычислений заполняем два последних столбца таблицы. Так как $\Sigma(x'+1)m_x = \Sigma x' m_x + \Sigma m_x$, то $\Sigma_7 = \Sigma_5 + \Sigma_2$. В самом деле, $45 = (-72) + 117$.

§ 1.5. Медиана и мода

Наряду со средними величинами в качестве описательных характеристик вариационного ряда применяют медиану и моду.

Медианой ($\tilde{M}_e$) называют значение признака, приходящееся на середину ранжированного ряда наблюдений.

Пусть проведено нечетное число наблюдений, т. е. $n = 2q - 1$, и результаты наблюдений проранжированы и выписаны в следующий ряд: $x_{(1)}, x_{(2)}, \ldots, x_{(q-1)}, x_{(q)}, x_{(q+1)}, \ldots, x_{(n)}$. Здесь $x_{(i)}$ — значение признака, занявшее i-е порядковое место в ранжированном ряду. На середину ряда приходится значение $x_{(q)}$. Следовательно,

$$\tilde{M}_e = x_{(q)}$$

Если проведено четное число наблюдений, т. е. $n = 2q$, то на середину ранжированного ряда $x_{(1)}, \ldots, x_{(q)}, x_{(q+1)}, \ldots, x_{(n)}$ приходятся значения $x_{(q)}$ и $x_{(q+1)}$. В этом случае за медиану принимают среднюю арифметическую значений $x_{(q)}$ и $x_{(q+1)}$, т. е.

$$\tilde{M}_e = (x_{(q)} + x_{(q+1)})/2.$$

Покажем на примерах, как определяется медиана дискретного и интервального вариационных рядов.

Пример 1.12. Вычислить медиану дискретного вариационного ряда, заданного табл. 1.1.

Решение. Проведено четное число наблюдений: $n = 100 = 2 \cdot 50$, следовательно,

$$\tilde{M}_e = (x_{(50)} + x_{(51)})/2,$$

где $x_{(50)}$ и $x_{(51)}$ — значения, расположенные соответственно на 50-м и 51-м порядковых местах в ранжированном ряду наблюдений. Чтобы найти $x_{(50)}$ и $x_{(51)}$ по вариационному ряду, воспользуемся тем фактом, что накопленная частота — это число, показывающее, сколько наблюдалось значений признака, меньших данного. По ряду накопленных частот (см. табл. 1.2) видим, что наблюдалось 38 значений признака, меньших пяти, и 82 значения, меньших шести. Следовательно, $x_{(50)} = x_{(51)} = 5$, а медиана $\tilde{M}_e = (5+5)/2 = 5$ (тарифный разряд).

Пример 1.13. Вычислить медиану интервального вариационного ряда, заданного табл. 1.3.

Решение. Проведено нечетное число наблюдений $n = 117$. Так

как $117=2\cdot59-1$, то $\tilde{M}_e=x_{(59)}$, где $x_{(59)}$ — значение признака, расположенное на 59-м порядковом месте в ранжированном ряду наблюдений. Однако по интервальному ряду нельзя определить точно ни одно из наблюдений, в том числе и нужное. По ряду накопленных частот видим, что наблюдалось 23 значения признака, меньших 100, и 69 значений, меньших 110. Следовательно, медиана, равная $x_{(59)}$, принадлежит интервалу 100—110. Этот интервал называют *медианным*. Заметим, что интервал 100—110 (медианный) был первым интервалом, которому соответствовала накопленная частота, превышающая половину всех наблюдений ($69>117/2$).

В медианном интервале наблюдалось 46 значений признака. Допустим, что в этом интервале наблюдения распределены равномерно. Тогда на единицу частоты медианного интервала приходится длина, равная $(110-100)/46$, и $x_{(59)}$ можно определить, если к началу медианного интервала прибавить часть его длины, приходящуюся на разность между 59 и частотой, накопленной к началу медианного интервала, т. е. $\tilde{M}_e=100+(10/46)\,(59-23)\approx107{,}8$.

В общем случае медиана для интервального вариационного ряда определяется по формуле

$$\tilde{M}_e = a_e + h\left(n/2 - m_e^{\text{нак}}\right)/m_e \qquad (1.9)$$

или по следующей формуле, полученной из формулы (1.9) в результате деления числителя и знаменателя входящей в нее дроби на n:

$$\tilde{M}_e = a_e + h\left(0{,}5 - w_e^{\text{нак}}\right)/w_e, \qquad (1.10)$$

где a_e — начало медианного интервала, т. е. такого, которому соответствует первая из накопленных частот (накопленных частостей), равная или большая половине всех наблюдений ($\geqslant0{,}5$); $m_e^{\text{нак}}$ ($w_e^{\text{нак}}$) — частота (частость), накопленная к началу медианного интервала; $m_e(w_e)$ — частота (частость) медианного интервала.

Пример 1.14. Вычислить медиану (в мм) для интервального вариационного ряда, заданного табл. 1.4.

Р е ш е н и е. По ряду накопленных частот определяем первую накопленную частоту, равную или бóльшую $n/2=200/2=100$. Она равна 130. Поэтому медианным является интервал 6,75—6,77 и $a_e=6{,}75$; $h=0{,}02$; $m_e^{\text{нак}}=78$; $m_e=52$. Следовательно, $M_e=6{,}75+0{,}02\cdot\dfrac{100-78}{52}\approx6{,}758$.

Модой ($\tilde{M}_0$) называют такое значение признака, которое наблюдалось наибольшее число раз. Нахождение моды для дискретного вариационного ряда не требует каких-либо вычислений, так как ею является вариант, которому соответствует наибольшая частота.

В случае интервального вариационного ряда мода вычисляется по следующей формуле*:

$$M_o = a_0 + h\,(m_0 - m_0')/(2m_0 - m_0' - m_0'') \qquad (1.11)$$

или по тождественной формуле:

$$M_o = a_0 + h\,(w_0 - w_0')/(2w_0 - w_0' - w_0''), \qquad (1.12)$$

где a_0 — начало модального интервала, т. е. такого, которому соответствует наибольшая частота (частость); m_0 (w_0) — частота (частость) модального интервала; m_0' (w_0') — частота (частость) интервала, предшествующего модальному; m_0'' (w_0'') — частота (частость) интервала, следующего за модальным.

Пример 1.15. Вычислить моду (в %) для интервального вариационного ряда, заданного табл. 1.3.

Решение. Модальным является интервал от 100 до 110, так как ему соответствует наибольшая частота, равная 46. Поэтому, подставляя в формулу (1.11) $a_0 = 100$, $h = 10$, $m_0 = 46$, $m_0' = 15$, $m_0'' = 29$, получаем

$$M_o = 100 + 10\,(46 - 15)/(2 \cdot 46 - 15 - 29) = 106{,}2.$$

Моду используют в случаях, когда нужно ответить на вопрос, какой товар имеет наибольший спрос, каковы преобладающие в данный момент уровни производительности труда, себестоимости и т. д. Модальная производительность, себестоимость и т. д. помогают вскрыть ресурсы, имеющиеся в экономике.

§ 1.6. Показатели вариации

Средние величины, характеризуя вариационный ряд числом, не отражают изменчивости наблюдавшихся значений признака, т. е. вариацию. Простейшим показателем вариации является *вариационный размах* ($R_в$), равный разности между наибольшим и наименьшим вариантами, т. е.

$$R_в = x_{max} - x_{min}. \qquad (1.13)$$

Вариационный размах — приближенный показатель вариации, так как почти не зависит от изменения вариантов, а крайние варианты, которые используются для его вычисления, как правило, ненадежны.

---

* Вывод формулы можно найти в кн.: *Венецкий И. Г., Кильдишев Г. С.* Теория вероятностей и математическая статистика. М., 1975.

Более содержательными являются меры рассеяния наблюдений вокруг средних величин. Средняя арифметическая является основным видом средних, поэтому ограничимся рассмотрением мер рассеяния наблюдений вокруг средней арифметической.

Сумма отклонений результатов наблюдений x_1, x_2, ..., x_n от средней арифметической $\bar{x}$, т. е. $\sum_{i=1}^{n}(x_i-\bar{x})$, не может характеризовать вариацию наблюдений около средней арифметической. В силу свойства 1° эта сумма равна нулю. Берут или абсолютные величины, или квадраты разностей $(x_i-\bar{x})$. В результате получают различные показатели вариации.

Средним линейным отклонением (d) называют среднюю арифметическую абсолютных величин отклонений результатов наблюдений от их средней арифметической:

$$d=\sum_{i=1}^{n}|x_i-\bar{x}|/n. \qquad (1.14)$$

Эмпирической дисперсией (s^2) называют среднюю арифметическую квадратов отклонений результатов наблюдений от их средней арифметической:

$$s^2=\sum_{i=1}^{n}(x_i-\bar{x})^2/n. \qquad (1.15)$$

Если по результатам наблюдений построен вариационный ряд, то эмпирическая дисперсия

$$s^2=\sum_{x}(x-\bar{x})^2 m_x/\sum_{x} m_x=\sum_{x}(x-\bar{x})^2 w_x. \qquad (1.16)$$

Вместо эмпирической дисперсии в качестве меры рассеяния наблюдений вокруг средней арифметической часто используют эмпирическое среднее квадратическое отклонение, равное арифметическому значению корня квадратного из дисперсии и имеющее ту же размерность, что и значения признака.

Для вариационного ряда среднее квадратическое отклонение

$$s=\sqrt{\sum_{x}(x-\bar{x})^2 m_x/\sum_{x} m_x}=$$

$$=\sqrt{\sum_{x}(x-\bar{x})^2 w_x}, \qquad (1.17)$$

где x — вариант (если ряд дискретный) и центр интервала (если ряд интервальный); $m_x(w_x)$ — соответствующая частота (частость); $\bar{x}$ — средняя арифметическая.

Пример 1.16. Для вариационного ряда, заданного табл. 1.3, вычислить эмпирическую дисперсию и эмпирическое среднее квадратическое отклонение.

Решение. Эмпирическую дисперсию вычислим по формуле (1.16), учитывая, что взвешенная средняя арифметическая $\bar{x} = 108{,}8\%$ (см. пример 1.9): $s^2 = (85-108{,}8)^2 \cdot (8/117) + (95-108{,}8)^2 \times 15/117 + (105-108{,}8)^2 (46/117) + (115-108{,}8)^2 (29/117) + (125-108{,}8)^2 \times 13/117 + (135-108{,}8)^2 (3/117) + (145-108{,}8)^2 (3/117) \approx 158{,}71$.

Эмпирическое среднее квадратическое отклонение

$$s = \sqrt{158{,}71} \approx 12{,}6\%.$$

Для краткости величину s^2 часто будем называть просто дисперсией, не употребляя термина «эмпирическая». Однако при этом всегда следует помнить, что в этом случае дисперсия вычислена по результатам наблюдений на основании опытных данных, т. е. является эмпирической. Аналогичное замечание относится и к величине s.

Приведем свойство минимальности эмпирической дисперсии: *s^2 меньше взвешенной средней арифметической квадратов отклонений вариантов от любой постоянной величины, отличной от средней арифметической, т. е.*

$$s^2 < \sum_x (x-a)^2 m_x \Big/ \sum_x m_x,$$

если $a \neq \bar{x}$.

Доказательство. Найдем экстремум функции $f(a) = \Sigma (x-a)^2 m_x / \Sigma m_x$. Для этого решим уравнение $f'(a) = 0$. Имеем:

$$f'_{(a)} = -2 \sum_x (x-a)\, m_x \Big/ \sum_x m_x = 0;$$

$$\sum_x (x-a)\, m_x = 0; \quad a = \sum_x x m_x \Big/ \sum_x m_x = \bar{x}.$$

Так как $f''(a) = 2\Sigma m_x / \Sigma m_x > 0$, то функция $f(a)$ имеет в точке $a = \bar{x}$ минимум. □

Можно показать, что среднее линейное отклонение не обладает свойством минимальности. Поэтому наиболее употребительными мерами рассеяния наблюдений вокруг средней арифметической являются эмпирическая

дисперсия и эмпирическое среднее квадратическое отклонение.

Итальянский статистик Коррадо Джинни предложил в качестве показателя вариации использовать величину

$$\Delta^2 = {}^1/_{n^2} \sum_{i,j=1}^{n} (x_i - x_j)^2,$$

где $x_1, x_2, \ldots, x_n$ — ряд наблюдений. Особенность этого показателя состоит в том, что он зависит только от разностей между наблюдениями и измеряет как бы «внутреннюю изменчивость» значений признака, а не их рассеяние вокруг какой-либо точки. Можно показать, что $s^2 = \Delta^2/2$, т. е. s^2, являясь мерой рассеяния значений признака вокруг средней арифметической, характеризует также и внутреннюю их изменчивость.

§ 1.7. Свойства эмпирической дисперсии

Рассмотрим основные свойства эмпирической дисперсии, знание которых позволит упростить ее вычисление.

1°. *Дисперсия постоянной величины равна нулю.*

Доказательство этого свойства очевидно вытекает из того, что дисперсия является показателем рассеяния наблюдений вокруг средней арифметической, а средняя арифметическая постоянной равна этой постоянной.

2°. *Если все результаты наблюдений уменьшить (увеличить) на одно и то же число c, то дисперсия не изменится*.*

Доказательство. Если все варианты уменьшить на число c, то в соответствии со свойством 2° средней арифметической средняя для измененного вариационного ряда равна $(\bar{x}-c)$, следовательно, его дисперсия

$$s^2_{x-c} = \frac{\sum_x [(x-c)-(\bar{x}-c)]^2 m_x}{\sum_x m_x} = \frac{\sum_x (x-\bar{x})^2 m_x}{\sum_x m_x} = s^2,$$

т. е. совпадает с дисперсией первоначального вариационного ряда. Аналогично можно показать, что $s^2_{x+c} = s^2$.

Доказанное свойство позволяет вычислять дисперсию не по данным вариантам, а по уменьшенным (увеличен-

* Доказательство свойств 2° и 3° проведем в предположении, что по результатам наблюдений построен вариационный ряд.

ным) на одно и то же число c, так как дисперсия, вычисленная для измененного ряда, равна первоначальной.

3°. *Если все результаты наблюдений уменьшить (увеличить) в одно и то же число k раз, то дисперсия уменьшится (увеличится) в k^2 раз.*

Доказательство. Если все варианты уменьшить в k раз, то, согласно свойству 3° средней арифметической, средняя для измененного вариационного ряда равна $\bar{x}/k$, следовательно, его дисперсия

$$s^2_{x/k} = \frac{\sum\limits_x (x/k - \bar{x}/k)^2 m_x}{\sum\limits_x m_x} = \frac{\sum\limits_x (1/k^2)(x - \bar{x})^2 m_x}{\sum\limits_x m_x} =$$

$$= \frac{1}{k^2} \frac{\sum\limits_x (x - \bar{x})^2 m_x}{\sum\limits_x m_x} = \frac{s^2}{k^2}. \quad \square$$

Аналогично можно показать, что $s^2_{xk} = k^2 s^2$.

Это свойство позволяет эмпирическую дисперсию вычислять не по данным вариантам, а по уменьшенным (увеличенным) в одно и то же число k раз. Если дисперсию, вычисленную для измененного ряда, увеличить (уменьшить) в k^2 раз, то получим дисперсию для первоначального вариационного ряда.

Следствие. *Если все варианты уменьшить (увеличить) в k раз, то среднее квадратическое отклонение уменьшится (увеличится) в число раз, равное k.*

Следствие очевидно вытекает из определения среднего квадратического отклонения.

Прежде чем рассматривать следующее свойство дисперсии, докажем теорему.

Теорема. *Эмпирическая дисперсия равна разности между средней арифметической квадратов наблюдений и квадратом средней арифметической, т. е.*

$$s^2 = \bar{x}^2 - (\bar{x})^2. \qquad (1.18)$$

Доказательство проведем для случая взвешенных средних арифметических, т. е. $\bar{x} = \sum\limits_x x m_x / \sum\limits_x m_x$, $\bar{x}^2 = \sum\limits_x x^2 m_x / \sum\limits_x m_x$.

Доказательство. Тождественно преобразуя выражения для дисперсии, имеем

$$s^2 = \frac{\sum\limits_x (x-\bar{x})^2 m_x}{\sum\limits_x m_x} = \frac{\sum\limits_x x^2 m_x - 2\sum\limits_x x\bar{x}m_x + \sum\limits_x (\bar{x})^2 m_x}{\sum\limits_x m_x} =$$

$$= \frac{\sum\limits_x x^2 m_x}{\sum\limits_x m_x} - \frac{2\bar{x}\sum\limits_x xm_x}{\sum\limits_x m_x} + \frac{(\bar{x})^2 \sum\limits_x m_x}{\sum\limits_x m_x} =$$

$$= \bar{x}^2 - 2(\bar{x})^2 + (\bar{x})^2 = \bar{x}^2 - (\bar{x})^2.$$

4°. *Если ряд наблюдений состоит из двух групп наблюдений, то дисперсия всего ряда равна сумме средней арифметической групповых дисперсий и средней арифметической квадратов отклонений групповых средних от средней всего ряда, причем при вычислении средних арифметических весами являются объемы групп.*

Пусть n_1 и n_2 — число наблюдений соответственно в 1-й и 2-й группах; $\bar{x}_1$, $\bar{x}_2$ — средние арифметические для 1-й и 2-й групп наблюдений; s_1^2, s_2^2 — дисперсии для 1-й и 2-й групп наблюдений; $\bar{x}$ и s^2 — средняя арифметическая и дисперсия для всего ряда n_1+n_2 наблюдений. Требуется доказать, что

$$s^2 = \frac{s_1^2 n_1 + s_2^2 n_2}{n_1 + n_2} + \frac{(\bar{x}_1 - \bar{x})^2 n_1 + (\bar{x}_2 - \bar{x})^2 n_2}{n_1 + n_2}.$$

Доказательство. Пусть $x_1, x_2, \ldots, x_{n_1+n_2}$ — ряд наблюдавшихся значений признака, причем к первой группе относятся наблюдения $x_1, x_2, \ldots, x_{n_1}$, а ко второй — наблюдения $x_{n_1+1}, x_{n_1+2}, \ldots, x_{n_1+n_2}$. Обозначим символом i порядковый номер наблюдения, попавшего в 1-ю группу, а через j — порядковый номер наблюдения, попавшего во 2-ю группу. На основании теоремы о дисперсии имеем $s_1^2 = \overline{x_i^2} - (\bar{x}_1)^2$, $s_2^2 = \overline{x_j^2} - (\bar{x}_2)^2$. Следовательно, первое слагаемое имеет вид

$$\frac{s_1^2 n_1 + s_2^2 n_2}{n_1 + n_2} = \frac{\overline{x_i^2} n_1 + \overline{x_j^2} n_2 - (\bar{x}_1)^2 n_1 - (\bar{x}_2)^2 n_2}{n_1 + n_2} =$$

$$= \left(\sum_{i=1}^{n_1} x_i^2 + \sum_{j=n_1+1}^{n_1+n_2} x_j^2 - (\bar{x}_1)^2 n_1 - (\bar{x}_2)^2 n_2 \right) / (n_1 + n_2) =$$

$$= \left(\sum_{q=1}^{n_1+n_2} x_q^2 - (\bar{x}_1)^2 n_1 - (\bar{x}_2)^2 n_2 \right) / (n_1 + n_2).$$

В соответствии со свойством 4° средней арифметической можно записать $\bar{x}(n_1+n_2) = \bar{x}_1 n_1 + \bar{x}_2 n_2$. Учитывая последнее равенство, преобразуем второе слагаемое:

$$\frac{(\bar{x}_1 - \bar{x})^2 n_1 + (\bar{x}_2 - \bar{x})^2 n_2}{n_1 + n_2} =$$

$$= \frac{(\bar{x}_1)^2 n_1 - 2\bar{x}_1 \bar{x} n_1 + (\bar{x})^2 n_1 + (\bar{x}_2)^2 n_2 - 2\bar{x}_2 \bar{x} n_2 + (\bar{x})^2 n_2}{n_1 + n_2} =$$

$$= \frac{(\bar{x}_1)^2 n_1 + (\bar{x}_2)^2 n_2 + (\bar{x})^2 (n_1 + n_2) - 2\bar{x}(\bar{x}_1 n_1 + \bar{x}_2 n_2)}{n_1 + n_2} =$$

$$= \frac{(\bar{x}_1)^2 n_1 + (\bar{x}_2)^2 n_2 - (\bar{x})^2 (n_1 + n_2)}{n_1 + n_2}.$$

Используя найденные выражения для слагаемых, получаем

$$\frac{s_1^2 n_1 + s_2^2 n_2}{n_1 + n_2} + \frac{(\bar{x}_1 - \bar{x})^2 n_1 + (\bar{x}_2 - \bar{x})^2 n_2}{n_1 + n_2} =$$

$$= \frac{\sum_{q=1}^{n_1+n_2} x_q^2 - (\bar{x}_1)^2 n_1 - (\bar{x}_2)^2 n_2 + (\bar{x}_1)^2 n_1 + (\bar{x}_2)^2 n_2 - (\bar{x})^2 (n_1+n_2)}{n_1 + n_2} =$$

$$= \sum_{q=1}^{n_1+n_2} x_q^2 / (n_1 + n_2) - (\bar{x})^2 = \overline{x^2} - (\bar{x})^2 = s^2. \quad \square$$

Свойство 4° можно обобщить на случай, когда ряд наблюдений состоит из любого количества $k \geqslant 2$ групп наблюдений. Введем понятия межгрупповой и внутригрупповой дисперсий.

Если ряд наблюдений состоит из k групп наблюдений, то *межгрупповой дисперсией* (δ^2) называют среднюю арифметическую квадратов отклонений групповых средних $\bar{x}_i$ от средней всего ряда наблюдений $\bar{x}$, причем весами являются объемы групп n_i, т. е.

$$\delta^2 = \sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2 n_i / \sum_{i=1}^{k} n_i.$$

Средней групповых дисперсий или внутригрупповой дисперсией $\overline{s_i^2}$ называют среднюю арифметическую груп-

повых дисперсий s_i^2, причем весами являются объемы групп n_i:

$$\overline{s_i^2} = \sum_{i=1}^{k} s_i^2 n_i / \sum_{i=1}^{k} n_i.$$

Следствие (свойства 4°). *Если ряд наблюдений состоит из k групп наблюдений, то дисперсия всего ряда s^2 равна сумме внутригрупповой и межгрупповой дисперсий, т. е.*

$$\overline{s^2} = \overline{s_i^2} + \delta^2.$$

Вычисление дисперсии вариационного ряда непосредственно по формуле (1.16) приводит к громоздким расчетам, если числовые значения вариантов и соответствующие им частоты велики. Поэтому часто дисперсию вычисляют не по первоначальным вариантам x, а по вариантам $x' = (x - c)/k$. Зная $s_{x'}^2$ (дисперсию для измененного ряда), легко вычислить дисперсию s^2 для первоначального ряда:

$$s^2 = s_{x'}^2 \cdot k^2. \qquad (1.19)$$

Действительно, принимая во внимание свойства 3° и 2° дисперсии, получаем

$$s_{x'}^2 = s_{(x-c)/k}^2 = (1/k^2)\, s_{x-c}^2 = (1/k^2)\, s^2,$$

откуда следует, что $s^2 = s_{x'}^2 \cdot k^2$.

Требования к c и k предъявляют те же, что и в упрощенном способе вычисления средней арифметической.

Пример 1.17. Вычислить дисперсию для распределения рабочих

Таблица 1.6

Выработка в отчетном году в процентах к предыдущему (интервалы)	Количество рабочих m_x	Середина интервала x	$x' = \frac{x-c}{k}$	$x'm_x$	$(x')^2 m_x$	$x'+1$	$(x'+1)\,m_x$	$(x'+1)^2 m_x$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
80—90	8	85	—3	—24	72	—2	—16	32
90—100	15	95	—2	—30	60	—1	—15	15
100—110	46	105	—1	—46	46	0	0	0
110—120	29	115	0	0	0	1	29	29
120—130	13	125	1	13	13	2	26	52
130—140	3	135	2	6	12	3	9	27
140—150	3	145	3	9	27	4	12	48
Σ	117	—	—	—72	230	—	—	203

по величине выработки в отчетном году в процентах к предыдущему году по данным, приведенным в табл. 1.3, используя предложенный способ.

Решение. Примем $c=115$, $k=10$. Результаты вычислений расположим в табл. 1.6. В соответствии с формулой (1.18) имеем

$$s_{x'}^2 = (\overline{x'})^2 - (\overline{x'})^2 = (230/117) - (-72/117)^2 \approx 1{,}5871.$$

По формуле (1.19) находим $s^2 = 1{,}5871 \cdot 10^2 = 158{,}71$. Для проверки хода вычислений заполняем три последних столбца таблицы. Так как

$$\Sigma (x'+1)^2 m_x = \Sigma (x')^2 m_x + 2 \Sigma x' m_x + \Sigma m_x,$$

то

$$\Sigma_3 = \Sigma_0 + 2\Sigma_5 + \Sigma_2.$$

Действительно, $203 = 230 + 2 \cdot (-72) + 117$.

§ 1.8. Эмпирические центральные и начальные моменты

Средняя арифметическая и дисперсия вариационного ряда являются частными случаями более общего понятия о моментах вариационного ряда.

Эмпирическим начальным моментом ($\tilde{\nu}_q$) порядка q называют взвешенную среднюю арифметическую q-х степеней вариантов, т. е.

$$\tilde{\nu}_q = \overline{x^q} = \sum_x x^q m_x / \sum_x m_x. \qquad (1.20)$$

Эмпирический начальный момент нулевого порядка

$$\tilde{\nu}_0 = \overline{x^0} = \sum_x x^0 m_x / \sum_x m_x = 1.$$

Эмпирический начальный момент первого порядка $\tilde{\nu}_1 = \overline{x}$.

Эмпирический начальный момент второго порядка

$$\tilde{\nu}_2 = \overline{x^2} = \sum_x x^2 m_x / \sum_x m_x \text{ и т. д.}$$

Эмпирическим центральным моментом ($\tilde{\mu}_q$) порядка q называют взвешенную среднюю арифметическую q-х степеней отклонений вариантов от их средней арифметической, т. е.

$$\tilde{\mu}_q = \overline{(x-\overline{x})^q} = \sum_x (x-\overline{x})^q m_x / \sum_x m_x. \qquad (1.21)$$

Эмпирический центральный момент нулевого порядка

$$\tilde{\mu}_0 = \overline{(x-\bar{x})^0} = \sum_x (x-\bar{x})^0 m_x / \sum_x m_x = 1.$$

Эмпирический центральный момент первого порядка

$$\tilde{\mu}_1 = \overline{(x-\bar{x})^1} = \sum_x (x-\bar{x})\, m_x / \sum_x m_x = 0$$

(в силу свойства 1° средней арифметической).

Эмпирический центральный момент второго порядка

$$\tilde{\mu}_2 = \overline{(x-\bar{x})^2} = \sum_x (x-\bar{x})^2 m_x / \sum_x m_x = s^2*.$$

Используя формулу бинома Ньютона, разложим в ряд выражение для центрального момента q-го порядка:

$$\tilde{\mu}_q = \overline{(x-\bar{x})^q} =$$

$$= \overline{x^q - C_q^1 x^{q-1}\bar{x} + C_q^2 x^{q-2}(\bar{x})^2 - \ldots + (-1)^q (\bar{x})^q} =$$

$$= \overline{x^q} - \overline{C_q^1 x^{q-1}\bar{x}} + \overline{C_q^2 x^{q-2}(\bar{x})^2} - \quad \ldots + \overline{(-1)^q (\bar{x})^q} =$$

$$= \overline{x^q} - C_q^1 \bar{x}\,\overline{x^{q-1}} + C_q^2 (\bar{x})^2 \overline{x^{q-2}} - \ldots + (-1)^q (\bar{x})^q =$$

$$= \tilde{\nu}_q - C_q^1 \tilde{\nu}_1 \tilde{\nu}_{q-1} + C_q^2 \tilde{\nu}_1^2 \tilde{\nu}_{q-2} - \ldots + (-1)^q \tilde{\nu}_1^q.$$

В проведенных тождественных преобразованиях использованы свойства 5° и 3° средней арифметической; $C_q^p = q!/(p!(q-p)!)$ — число сочетаний из q элементов по p элементов ($p \leqslant q$).

Итак, центральный момент q-го порядка выражается через начальные моменты следующим образом:

$$\tilde{\mu}_q = \tilde{\nu}_q - C_q^1 \tilde{\nu}_1 \tilde{\nu}_{q-1} + C_q^2 \tilde{\nu}_1^2 \tilde{\nu}_{q-2} - C_q^3 \tilde{\nu}_1^3 \tilde{\nu}_{q-3} + \\ + \ldots + (-1)^q \tilde{\nu}_1^q. \qquad (1.22)$$

Полагая $q = 0, 1, 2, \ldots$, можно получить выражения центральных моментов различных порядков через начальные моменты:

$$\tilde{\mu}_0 = \tilde{\nu}_0 = 1, \quad \tilde{\mu}_1 = \tilde{\nu}_1 - \tilde{\nu}_1 = 0;$$

---

* В дальнейшем для краткости величину $\tilde{\mu}_q(\tilde{\nu}_q)$ часто будем называть просто центральным моментом (начальным моментом), не употребляя термин «эмпирический».

$$\tilde{\mu}_2 = \tilde{\nu}_2 - 2\tilde{\nu}_1\tilde{\nu}_1 + \tilde{\nu}_1^2 = \tilde{\nu}_2 - \tilde{\nu}_1^2; \quad (1.23)$$

$$\tilde{\mu}_3 = \tilde{\nu}_3 - 3\tilde{\nu}_1\tilde{\nu}_2 + 3\tilde{\nu}_1^2\tilde{\nu}_1 - \tilde{\nu}_1^3 = \tilde{\nu}_3 - 3\tilde{\nu}_1\tilde{\nu}_2 + 2\tilde{\nu}_1^3; \quad (1.24)$$

$$\tilde{\mu}_4 = \tilde{\nu}_4 - 4\tilde{\nu}_1\tilde{\nu}_3 + 6\tilde{\nu}_1^2\tilde{\nu}_2 - 4\tilde{\nu}_1^3\tilde{\nu}_1 + \tilde{\nu}_1^4 =$$

$$= \tilde{\nu}_4 - 4\tilde{\nu}_1\tilde{\nu}_3 + 6\tilde{\nu}_1^2\tilde{\nu}_2 - 3\tilde{\nu}_1^4 \quad (1.25)$$

и т. д.

Заметим, что формула (1.23) для центрального момента второго порядка, как и следовало ожидать, аналогична формуле (1.18) для дисперсии.

Рассмотрим свойства центральных моментов, которые позволят значительно упростить их вычисление.

1°. *Если все варианты уменьшить (увеличить) на одно и то же число c, то центральный момент q-го порядка не изменится.*

Доказательство. Если все варианты уменьшить на число c, то средняя арифметическая для измененного ряда равна $\bar{x}-c$, поэтому центральный момент q-го порядка

$$\tilde{\mu}_{q_{x-c}} = \sum_x [(x-c) - (\bar{x}-c)]^q m_x / \sum_x m_x =$$

$$= \sum_x (x-\bar{x})^q m_x / \sum_x m_x = \tilde{\mu}_q. \quad \square$$

Аналогично можно показать, что $\tilde{\mu}_{q_{x+c}} = \tilde{\mu}_q$.

2°. *Если все варианты уменьшить (увеличить) в одно и то же число k раз, то центральный момент q-го порядка уменьшится (увеличится) в k^q раз.*

Доказательство. Если все варианты уменьшить в одно и то же число k раз, то средняя арифметическая для измененного вариационного ряда равна $\bar{x}/k$, поэтому центральный момент q-го порядка

$$\tilde{\mu}_{q_{x/k}} = \sum_x (x/k - \bar{x}/k)^q m_x / \sum_x m_x =$$

$$= (1/k^q) \sum_x (x-\bar{x})^q m_x / \sum_x m_x = \tilde{\mu}_q / k^q. \quad \square$$

Аналогично можно показать, что $\tilde{\mu}_{q_{x \cdot k}} = \tilde{\mu}_q \cdot k^q$.

Для облегчения расчетов центральные моменты вычисляют не по первоначальным вариантам x, а по вариантам $x' = (x - c)/k$. Зная μ'_q (центральный момент q-го порядка для измененного ряда), легко вычислить центральный момент q-го порядка $\tilde{\mu}_q$ для первоначального ряда:

$$\tilde{\mu}_q = \tilde{\mu}'_q \cdot k^q . \qquad (1.26)$$

Действительно, принимая во внимание свойства центрального момента, получаем

$$\tilde{\mu}'_q = \tilde{\mu}_{q\,(x-c)/k} = (1/k^q)\,\tilde{\mu}_{q\,x-c} = (1/k^q)\,\tilde{\mu}_q,$$

откуда следует, что $\tilde{\mu}_q = \tilde{\mu}'_q \cdot k^q$.

Пример 1.18. Вычислить центральные моменты второго, третьего и четвертого порядков для распределения рабочих по величине выработки в отчетном году в процентах к предыдущему по данным, приведенным в табл. 1.3.

Решение. Результаты вычислений расположим в табл. 1.7.

Таблица 1.7

Выработка на одного рабочего в отчетном году в процентах к предыдущему (интервалы)	Количество рабочих m_x	Середина интервала x	$x' = \frac{x-c}{k}$	$x'm_x$	$(x')^2 m_x$	$(x')^3 m_x$	$(x')^4 m_x$	$x'+1$	$(x'+1)\,m_x$	$(x'+1)^2 m_x$	$(x'+1)^3 m_x$	$(x'+1)^4 m_x$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
80—90	8	85	—3	—24	72	—216	+648	—2	—16	32	—64	128
90—100	15	95	—2	—30	60	—120	+240	—1	—15	15	—15	15
100—110	46	105	—1	—46	46	—46	46	0	0	0	0	0
110—120	29	115	0	0	0	0	0	1	29	29	29	29
120—130	13	125	1	13	13	13	13	2	26	52	104	208
130—140	3	135	2	6	12	24	48	3	9	27	81	243
140—150	3	145	3	9	27	81	243	4	12	48	192	768
Σ	117			—72	230	—264	1238			203	327	1391

В соответствии с определением начального момента получим следующие значения начальных моментов для вариантов x':

$$\tilde{\nu}'_1 = \sum x' m_x / \sum m_x = -72/117;\ \tilde{\nu}'_2 = \sum (x')^2 m_x / \sum m_x = 230/117;$$

$$\tilde{\nu}'_3 = \sum (x')^3 m_x / \sum m_x = -264/117;\ \tilde{\nu}'_4 = \sum (x')^4 m_x / \sum m_x = 1238/117.$$

Используя формулы (1.23)—(1.25), получаем следующие значения центральных моментов для вариантов x':

$$\tilde{\mu}_2' = \tilde{\nu}_2' - (\tilde{\nu}_1')^2 \approx 1{,}5871;$$

$$\tilde{\mu}_3' = \tilde{\nu}_3' - 3\tilde{\nu}_1'\tilde{\nu}_2' + 2(\tilde{\nu}_1')^3 \approx 0{,}90669;$$

$$\tilde{\mu}_4' = \tilde{\nu}_4' - 4\tilde{\nu}_1'\tilde{\nu}_3' + 6(\tilde{\nu}_1')^2\tilde{\nu}_2' - 3(\tilde{\nu}_1')^4 \approx 9{,}063416.$$

На основании формулы (1.26) получаем следующие значения центральных моментов для первоначального вариационного ряда:

$$\tilde{\mu}_2 = 1{,}5871 \cdot 10^2 = 158{,}71; \quad \tilde{\mu}_3 = 0{,}90669 \cdot 10^3 = 906{,}69;$$

$$\tilde{\mu}_4 = 9{,}063416 \cdot 10^4 = 90634{,}16.$$

Заметим, что дисперсия, вычисленная в примере 1.17, совпадает со значением центрального момента второго порядка.

Для проверки хода вычислений заполняем пять последних столбцов таблицы. Так как

$$\Sigma (x'+1)^3 m_x = \Sigma (x')^3 m_x + 3\Sigma (x')^2 m_x + 3\Sigma x' m_x + \Sigma m_x \text{ и}$$

$$\Sigma (x'+1)^4 m_x = \Sigma (x')^4 m_x + 4\Sigma (x')^3 m_x + 6\Sigma (x')^2 m_x + 4\Sigma x' m_x + \Sigma m_x, \text{ то}$$

$$\Sigma_{12} = \Sigma_7 + 3\Sigma_6 + 3\Sigma_5 + \Sigma_2 \text{ и } \Sigma_{13} = \Sigma_8 + 4\Sigma_7 + 6\Sigma_6 + 4\Sigma_5 + \Sigma_2.$$

Действительно, $327 = -264 + 3 \cdot 230 + 3(-72) + 117$ и $1391 = 1238 + 4(-264) + 6 \cdot 230 + 4 \cdot (-72) + 117$.

§ 1.9. Эмпирические асимметрия и эксцесс

Эмпирическим коэффициентом асимметрии $\tilde{A}$ называют отношение центрального момента третьего порядка к кубу среднего квадратического отклонения:

$$\tilde{A} = \frac{\tilde{\mu}_3}{s^3} = \frac{\tilde{\mu}_3}{\left(\sqrt{\tilde{\mu}_2}\right)^3} = \frac{\sum_x (x - \bar{x})^3 m_x}{(\Sigma m_x) s^3}. \qquad (1.27)$$

Если полигон вариационного ряда скошен, т. е. одна из его ветвей, начиная от вершины, зримо длиннее другой, то такой ряд называют *асимметричным*. Из формулы (1.27) следует, что если в варнационном ряду преобладают варианты, меньшие $\bar{x}$, то эмпирический коэффициент асимметрии отрицателен; говорят, что в этом случае имеет место *левосторонняя асимметрия*. Если же в вариационном ряду преобладают варианты, большие $\bar{x}$, то эм-

пирический коэффициент асимметрии положителен; в этом случае имеет место *правосторонняя асимметрия.* При левосторонней асимметрии левая ветвь полигона длиннее правой. При правосторонней, более длинной является правая ветвь.

Эмпирический коэффициент асимметрии не имеет ни верхней, ни нижней границы, что снижает его ценность как меры асимметрии. Практически коэффициент асимметрии редко бывает особенно велик, а для умеренно асимметричных рядов он обычно меньше единицы.

Эмпирическим эксцессом или *коэффициентом крутости* $\tilde{E}$ называют уменьшенное на 3 единицы отношение центрального момента четвертого порядка к четвертой степени среднего квадратического отклонения:

$$\tilde{E} = \tilde{\mu}_4/s^4 - 3 = \tilde{\mu}_4/\tilde{\mu}_2^2 - 3. \qquad (1.28)$$

За стандартное значение эксцесса принимают нуль-эксцесс так называемой нормальной кривой (см. рис. 3.15). Кривые, у которых эксцесс отрицательный, по сравнению с нормальной менее крутые, имеют более плоскую вершину и называются «плосковершинными» Кривые с положительным эксцессом более крутые по сравнению с нормальной кривой, имеют более острую вершину и называются «островершинными».

Пример 1.19. По данным, приведенным в табл. 1.3, вычислить эмпирический коэффициент асимметрии и эксцесс для распределения рабочих по величине выработки в отчетном году в процентах к предыдущему.

Р е ш е н и е. Центральные моменты второго, третьего и четвертого порядка, необходимые для вычисления эмпирической асимметрии и эксцесса, вычислены в примере 1.18. ($\tilde{\mu}_2 = 158{,}71$, $\tilde{\mu}_3 = 906{,}69$, $\tilde{\mu}_4 = 90634{,}16$). Коэффициент асимметрии $\tilde{A} = 906{,}69 : (\sqrt{158{,}71})^3 = 0{,}45$; эксцесс $\tilde{E} = 90634{,}16/158{,}71^2 - 3 = 0{,}60$. Коэффициент асимметрии оказался положительным. Это свидетельствует о правосторонней асимметрии данного распределения; действительно, правая ветвь этого распределения (см. рис. 1.2) зримо длиннее левой.

Эксцесс оказался также положительным.

Пример 1.20. По данным, приведенным в табл. 1.4, вычислить средний размер диаметра валика после шлифовки, эмпирическое среднее квадратическое отклонение, коэффициент асимметрии и эксцесс.

Р е ш е н и е. Результаты вычислений и проверку хода вычислений поместим в табл. 1.8.

Таблица 1.8

Диаметр валика после шлифовки (интервалы)	Количество валиков m_x	Середина интервала x	$x' = \frac{x - A}{h}$	$x' m_x$	$(x')^2 m_x$	$(x')^3 m_x$	$(x')^4 m_x$	$x'+1$	$(x'+1) m_x$	$(x'+1)^2 m_x$	$(x'+1)^3 m_x$	$(x'+1)^4 m_x$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
6,67—6,69	2	6,68	—4	—8	32	—128	512	—3	—6	18	—54	162
6,69—6,71	15	6,70	—3	—45	135	—405	1215	—2	—30	60	—120	240
6,71—6,73	17	6,72	—2	—34	68	—136	272	—1	—17	17	—17	17
6,73—6,75	44	6,74	—1	—44	44	—44	44	0	0	0	0	0
6,75—6,77	52	6,76	0	0	0	0	0	1	52	52	52	52
6,77—6,79	44	6,78	1	44	44	44	44	2	88	176	352	704
6,79—6,81	14	6,80	2	28	56	112	224	3	42	126	378	1134
6,81—6,83	11	6,82	3	33	99	297	891	4	44	176	704	2816
6,83—6,85	1	6,84	4	4	16	64	256	5	5	25	125	625
Σ	200			—22	494	—196	5458		178	650	1420	5750

$c = 6{,}76, \qquad k = h = 0{,}02$

Проверка	Характеристика для x'	Характеристика для x
$\Sigma_{10} = \Sigma_5 + \Sigma_2$ $178 = -22 + 200$	$\bar{x}' = \Sigma_5/\Sigma_2 =$ $= -22/200 = -0{,}11$	$\bar{x} = \bar{x}' \cdot k + c$ $\bar{x} = -0{,}11 \cdot 0{,}02 + 6{,}76 = 6{,}7578$ (мм)
$\Sigma_{11} = \Sigma_6 + 2\Sigma_5 + \Sigma_2$ $650 = 494 +$ $+ 2(-22) + 200$	$s_{x'}^2 = (\overline{x'})^2 - (\overline{x'})^2,$ $s_{x'}^2 = \Sigma_6/\Sigma_2 - (-0{,}11)^2 = \frac{494}{200} - 0{,}0121 =$ $= 2{,}4579$	$s^2 = s_{x'}^2 \cdot k^2$ $s^2 = 2{,}4579 \cdot 0{,}02^2 = 0{,}9832 \cdot 10^{-3}$ мм^2 $s = \sqrt{s^2} = 0{,}0314$ мм
$\Sigma_{12} = \Sigma_7 + 3\cdot\Sigma_6 +$ $+ 3\Sigma_5 + \Sigma_2$ $42 = -196 + 3\cdot 494 +$ $+ 3(-22) + 200$	$\tilde{\mu}_3' = \tilde{\nu}_3' - 3\tilde{\nu}_1'\tilde{\nu}_2' + 2(\tilde{\nu}_1')^3, \qquad \tilde{\mu}_3' = \Sigma_7/\Sigma_2 -$ $- 3\Sigma_5/\Sigma_2\,\Sigma_6/\Sigma_2 + 2(\Sigma_5/\Sigma_2)^3,$ $\mu_3' = -196/200 - 3(-22)/200 \cdot 494/200 +$ $+ 2(-22/200) = -0{,}1675$	$\tilde{\mu}_3 = \tilde{\mu}_3' \cdot k^3$ $\tilde{\mu}_3 = -0{,}1675 \cdot 0{,}02^3 = 0{,}1340 \cdot 10^{-4}$ $\tilde{A} = \tilde{\mu}_3/s^3 = -0{,}04$
$\Sigma_{13} = \Sigma_8 + 4\Sigma_7 + 6\Sigma_6 +$ $+ 4\Sigma_5 + \Sigma_2$ $5750 = 3458 +$ $+ 4(-196) + 6\cdot 494 +$ $+ 4(-22) + 200$	$\tilde{\mu}_4' = \tilde{\nu}_4' - 4\tilde{\nu}_1'\tilde{\nu}_3' + 6(\tilde{\nu}_1')^2\tilde{\nu}_2' - 3(\tilde{\nu}_1')^4.$ $\tilde{\mu}_4' = \Sigma_8/\Sigma_2 - 4\Sigma_5/\Sigma_2\,\Sigma_7/\Sigma_2 + 6(\Sigma_5/\Sigma_2)^2\,\Sigma_6/\Sigma_2 -$ $- 3(\Sigma_5/\Sigma_2)^4$ $\mu_4' = 3458/200 - 4(-22)/200 \cdot (-196)/200 +$ $+ 6(-22/200)^2 \cdot 494/200 - 3(-22/200)^4 =$ $= 17{,}0376.$	$\tilde{\mu}_4 = \tilde{\mu}_4' \cdot k^4$ $\tilde{\mu}_4 = 17{,}0376 \cdot 0{,}02^4 = 0{,}2726 \cdot 10^{-5}$ $\tilde{E} = \tilde{\mu}_4/s^4 - 3 = -0{,}18$

Глава 2

Основные понятия теории вероятностей

§ 2.1. Классификация событий

Одним из основных понятий теории вероятностей является понятие события. Под *событием* понимают любой факт, который может произойти в результате опыта, или испытания. Под *опытом*, или *испытанием*, понимают осуществление определенного комплекса условий.

Простейшим примером события является результат подбрасывания монеты. В этом примере опытом является подбрасывание монеты, а его исходы — событиями. В данном опыте рассматриваются два события: появление цифры и появление «герба» Событием можно, например, считать выход со станка деталей определенного размера или деталей, размеры которых лежат внутри определенного интервала. Появление бракованных изделий также событие. Можно привести бесчисленное множество подобных примеров. События обозначают заглавными буквами латинского алфавита A, B, C, ...

Различают события *совместные* и *несовместные*. События называют совместными, если наступление одного из них сопровождается наступлением других в одном и том же испытании. В противном случае события являются несовместными. Рассмотрим следующий пример. Завод выпускает приборы трех типоразмеров: I, II и III. Пусть A — изготовленный прибор относится к I типоразмеру; B — изготовленный прибор относится ко II типоразмеру; C — изготовленный прибор относится к III типоразмеру. Очевидно, что эти события являются несовместными.

Техническим контролем изделия могут быть забракованы по нескольким признакам: выход размеров за границы технического допуска, наличие механических пов-

реждений, непроваренность сварного шва и т. д. Одному изделию могут быть присущи все указанные виды дефектов. Поэтому эти виды дефектов являются совместными событиями.

Событие называют *достоверным*, если оно обязательно произойдет в условиях данного опыта.

Событие называют *невозможным*, если оно не может произойти в условиях данного опыта.

Если, например, двигатель исправен, нормально функционирует система топливоподачи и источник энергии (аккумулятор) находится в рабочем состоянии, то при включении зажигания и стартера вращение вала двигателя автомобиля — событие достоверное. При выходе из строя хотя бы одной системы топливоподачи вращение вала двигателя становится событием невозможным.

Событие называют *возможным* или *случайным*, если в результате опыта оно может появиться, но может и не появиться.

Примером случайного события могут служить выявление дефектного изделия при контроле партии готовой продукции, несоответствие размера обрабатываемого изделия заданному, отказ одного из звеньев автоматизированной системы управления.

События называют *равновозможными*, если по условиям испытания ни одно из этих событий не является объективно более возможным, чем другое.

Приведем следующий пример. Пусть магазину поставляют электролампы (причем в равных количествах) несколько заводов-изготовителей. События, состоящие в покупке лампы любого из этих заводов, равновозможны.

Важным понятием является *полная группа событий*. Несколько событий в данном опыте образуют полную группу, если в результате опыта обязательно появится хотя бы одно из них. Например, если в механическом цехе весь парк станков разбит на следующие группы (события): A — станку требуется наладка через 7 ч работы; B — 10 ч; C — 15 ч; D — 20 ч; E — 21 ч и более, то в данном случае события A, B, C, D, E образуют полную группу.

Введем понятие противоположного или дополнительного события. Под *противоположным* событием A понимают событие, которое обязательно должно произойти, если не наступило некоторое событие A. Противополож-

ные события несовместны и единственно возможны. Они образуют полную группу событий. Так, например, если партия изготовленных изделий состоит из годных и бракованных, то при извлечении одного изделия оно может оказаться либо годным — событие A, либо бракованным — событие $\bar{A}$.

События, образующие полную группу событий и являющиеся несовместными и равновозможными, называют *случаями* или *шансами*.

Если наблюдающиеся в опыте события равновозможны, несовместны и образуют полную группу, то говорят, что опыт сводится к *схеме случаев* или к *схеме урн*. (Эти термины впервые появились в работах, посвященных исследованию правил азартных игр при зарождении науки о вероятностях.)

§ 2.2. Классическое определение вероятности события

Для количественного сравнения между собой событий по степени возможности их появления вводится определенная мера, которая называется вероятностью события.

Вероятностью события называется численная мера степени объективной возможности появления этого события.

Рассмотрим следующее понятие. Случай называют *благоприятным* или *благоприятствующим* некоторому событию, если появление этого случая влечет за собой появление данного события. Пусть, например, производится опыт — подбрасывание кубика, грани которого пронумерованы цифрами от 1 до 6. Рассмотрим событие A — появление четной цифры. Случаи появления чисел 2, 4 и 6 благоприятствуют событию A. Приведем еще один пример. Пусть в ящике содержатся годные изделия и дефектные, забракованные по нескольким видам дефектов. Рассмотрим событие, заключающееся в появлении бракованного изделия при вынимании его наугад из ящика. Этому событию будут благоприятствовать все изделия, имеющие дефекты.

Если опыт сводится к схеме случаев, то вероятность появления события A в данном опыте вычисляется как отношение числа случаев M, благоприятствующих появлению этого события, к общему числу случаев N. Это и есть *классическое определение вероятности события*. Та-

ким образом, вероятность события A

$$P(A) = M/N. \qquad (2.1)$$

Впервые определение вероятности события было дано в XVIII в. Лапласом. Из формулы (2.1) следует, что вероятность события может меняться в пределах от 0 до 1 в зависимости от того, какую долю составляет благоприятствующее число случаев от общего числа случаев, т. е. $0 \leqslant M \leqslant N, 0 \leqslant P(A) \leqslant 1$. Исходя из приведенного определения вероятности можно убедиться в следующем:

1°. *Вероятность достоверного события равна единице, так как для того чтобы событие обязательно произошло, все случаи в опыте должны ему благоприятствовать, т. е.* $M = N$, *тогда* $P(A) = 1$.

2°. *Вероятность невозможного события равна нулю, так как ни один случай в опыте не благоприятствует такому событию, т. е.* $M = 0$, *тогда* $P(A) = 0$.

3°. *Вероятность наступления событий, образующих полную группу, равна единице, так как появление хотя бы одного из них в результате опыта является событием достоверным.*

4°. *Вероятность наступления противоположного события* $\bar{A}$ *равна разности между единицей и вероятностью наступления события* A:

$$P(\bar{A}) = 1 - P(A), \qquad (2.2)$$

так как $P(\bar{A}) = (N-M)/n = 1 - M/N$, *где* $N - M$ — *число случаев, благоприятствующих появлению противоположного события* $\bar{A}$.

Важным достоинством классического определения вероятности события является то, что с его помощью вероятность события можно определить, не прибегая к опыту, а исходя из логических рассуждений. Приведем несколько примеров непосредственного вычисления вероятностей.

Пример 2.1. В ящике находятся 10 бракованных и 15 годных деталей, которые тщательно перемешаны. Найти вероятность того, что наудачу извлеченная деталь годная.

Р е ш е н и е. Обозначим через A событие, состоящее в появлении годной детали. Общее число случаев $N = 25$, число случаев, благоприятных событию A, $M = 15$. Следовательно, $P(A) = 15/25 = 3/5$.

Пример 2.2. Условия предыдущей задачи сохраняются, только теперь из ящика последовательно извлекают три детали (без возвращения): а) чему равна вероятность того, что все три детали будут годными? б) какова вероятность того, что извлечены только две годные детали?

Р е ш е н и е. а) Обозначим через B событие, состоящее в появлении трех годных деталей. Общее число случаев равно числу способов, с помощью которых можно извлечь из ящика все детали по три (числу сочетаний из 25 элементов по три): $N=C_{25}^{3}$. Число случаев, благоприятствующих событию B, равно числу способов, позволяющих взять три детали из числа годных, т. е. числу сочетаний из 15 элементов по три: $N=C_{15}^{3}$. Искомая вероятность $P(B)=C_{15}^{3}/C_{25}^{3}=91/460$.

б) Обозначим через D событие, состоящее в появлении двух годных деталей. Общее число случаев останется прежним, т. е. $N=C_{25}^{3}$. Число случаев, благоприятствующих событию D, равно числу способов, с помощью которых можно одновременно взять две годные детали из 15 годных и одну бракованную из 10 бракованных, т. е. $M=C_{15}^{2}C_{10}^{1}$; $P(D)=C_{15}^{2}C_{10}^{1}/C_{25}^{3}=91/460$.

§ 2.3. Статистическое определение вероятности события

Формулу (2.1) используют для непосредственного вычисления вероятностей событий только тогда, когда опыт сводится к схеме случаев.

На практике часто классическое определение вероятности неприменимо по двум причинам: во-первых, классическое определение вероятности предполагает, что общее число случаев N должно быть конечно. На самом же деле оно зачастую неограниченно; во-вторых, часто невозможно представить исходы опыта в виде схемы случаев.

Давно было замечено, что частота появления событий, не сводящихся к схеме случаев, при многократно повторяющихся опытах имеет тенденцию стабилизироваться около какой-то постоянной величины. Это свидетельствует о том, что данные события также обладают определенной степенью объективной возможности появления в опыте, меру которой можно представить в виде относительной частоты (или частости).

Знаменитый швейцарский ученый Яков Бернулли привел математическое доказательство того, что при большом числе испытаний частость стремится воспроизвести вероятность и в пределе при большом числе опытов должна практически совпадать с ней. Это положение носит название *закона больших чисел* (этот закон будет рассмотрен в дальнейшем). Многие ученые проводили специальные опыты для проверки закона Бернулли, вычисляя частость появления событий, сводящихся к схеме

случаев, и сравнивая ее с вероятностью, вычисленной по классическому определению. Так, французский натуралист Бюффон и английский статистик Пирсон проводили опыты с бросанием монеты (подсчитывалось число случаев выпадения «герба»). Результаты их опытов приведены в табл. 2.1.

Таблица 2.1

Опыты	Вероятность	Частость	Отклонение частости от вероятности	Число испытаний
Опыт Бюффона	0,5000	0,5069	0,0069	4 040
Первый опыт Пирсона	0,5000	0,5016	0,0016	12 000
Второй опыт Пирсона	0,5000	0,5005	0,0005	24 000

Как следует из таблицы, уже 4040 испытаний достаточно для совпадения эмпирической частости и вероятности до сотого знака, а 24 000 испытаний дают практически полное совпадение частости и вероятности. Таким образом, можно считать эмпирически подтвержденным, что частость событий, сводящихся к схеме случаев, при большом числе произведенных опытов стремится к их вероятности. Это свойство случайных событий, известное под названием *устойчивости частот*, можно перенести и на другие типы случайных событий, не обладающих симметрией исходов и не составляющих полную группу. Правда, подсчет теоретической вероятности для таких событий представляет известные трудности.

Так возникло понятие *статистической* вероятности события, под которой понимается относительная частота появления события A в N произведенных опытах. Статистическая вероятность вычисляется на основании результатов опытов по следующей формуле *:

$$w(A) = M/N, \qquad (2.3)$$

где M — число появлений события A в N опытах.

В § 2.2 было показано, что вероятность достоверного события равна единице, а вероятность невозможного

* Статистическая вероятность события идентична понятию частости появления определенного значения признака, приведенному в гл. 1.

равна нулю. Однако на практике приходится сталкиваться с событиями, вероятности которых не точно равны единице или нулю, а весьма близки к этим значениям. Первые из этих событий называют *практически достоверными,* а вторые —*практически невозможными.* Если события имеют такие вероятности, то можно предсказать результат опыта, руководствуясь так называемым *принципом практической уверенности,* который заключается в следующем: *если вероятность некоторого события A в данном опыте весьма мала, то можно быть практически уверенным в том, что при однократном выполнении этого опыта событие A не произойдет.*

Вопрос о том, насколько мала или велика должна быть вероятность, решается в каждом отдельном случае в зависимости от важности исследуемого явления и не является предметом обсуждения в теории вероятностей.

§ 2.4. Понятия суммы и произведения событий

Непосредственный подсчет вероятностей по классическому или статистическому определению для решения многих практически важных задач достаточно труден. Поэтому возникла необходимость в разработке методов вычисления вероятностей событий, позволяющих по известным вероятностям одних событий определять вероятности наступления других событий. Эти методы связаны с применением основных теорем теории вероятностей: теоремы сложения и теоремы умножения вероятностей, которые могут быть строго доказаны только для событий, сводящихся к схеме случаев, однако их можно распространить и на события, не сводящиеся к этой схеме. В этом случае их принимают как аксиомы теории вероятностей. Прежде чем сформулировать и доказать основные теоремы, рассмотрим некоторые вспомогательные понятия, а именно понятия суммы и произведения событий. .

Суммой двух событий A и *B* называют событие *C*, состоящее в появлении события *A*, или события *B*, или обоих этих событий. Другими словами, можно сказать, что событие *C* состоит в появлении хотя бы одного из событий *A* и *B*. Например, если изготовленные изделия бракуют по двум видам дефектов: *A* — несоблюдение технического допуска на геометрические размеры и *B*— недостаточная твердость после закалки, то событие *C*

означает вообще наличие дефекта у проверяемого изделия (либо несоблюдение технического допуска, либо недостаточная твердость после закалки, либо оба вместе). Если события A и B несовместны, то появление этих событий вместе невозможно и сумма событий A и B заключается в появлении или события A, или события B.

Суммой нескольких событий называют событие, состоящее в появлении хотя бы одного из этих событий.

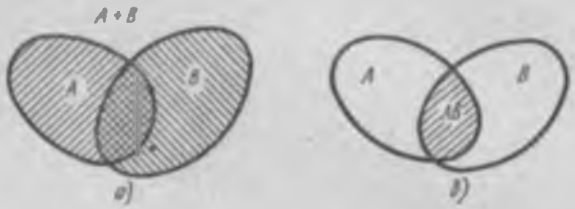

Рис. 2.1

Произведением двух событий A и B называют событие C, состоящее в совместном появлении события A и события B.

Например, если взять для контроля две детали из партии, состоящей из бракованных и годных деталей, и событие A означает, что первая взятая деталь годная; событие B—вторая взятая деталь также годная, то собы-

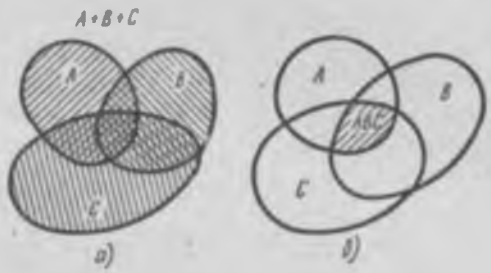

Рис. 2.2

тие $C=AB$ состоит в том, что обе детали годные.

Произведением нескольких событий называют событие, состоящее в совместном появлении всех этих событий.

Дадим геометрическую интерпретацию понятий суммы и произведения событий. Если обозначить через A попадание точки в область A, через B — попадание точки в область B, то $A+B$ — попадание в область, заштрихо-

ванную на рис. 2.1, *а*; AB—попадание в область, заштрихованную на рис. 2.1, *б*. На рис. 2.2, *а, б* дана геометрическая интерпретация суммы и произведения трех событий.

§ 2.5. Теорема сложения вероятностей

Предварительно рассмотрим конкретный пример. Пусть имеется партия гаек со следующим распределением размеров по диаметру: 120 гаек диаметром 24,90 мм; 40 гаек — 24,86 мм; 30 гаек — 24,85 мм; 10 гаек — 24,84 мм. Какова вероятность извлечь гайку диаметром либо 24,86, либо 24,85, либо 24,84 мм?

Общее количество случаев равно 200. Из них благоприятствуют поставленному условию 80. В соответствии с этим вероятность извлечения гайки диаметром от 24,86 до 24,84 мм $P(A)=80/200=2/5$. Теперь рассмотрим вероятность извлечения гаек каждого диаметра в отдельности. Для этого обозначим через A_1 событие, заключающееся в извлечении из данной совокупности гайки диаметром 24,90 мм, A_2 — гайки диаметром 24,86 мм, A_3 — гайки диаметром 24,85 мм, A_4 — гайки диаметром 24,84 мм. Тогда имеем: $P(A_1)=120/200=3/5$; $P(A_2)=40/200=1/5$; $P(A_3)=30/200=3/20$; $P(A_4)=10/200=1/20$.

По условию задачи необходимо определить вероятность извлечения гаек с диаметром либо 24,86, либо 24,85, либо 24,84 мм. Складывая вероятности появления гаек определенного диаметра, находим $P(A)=1/5+3/20+1/20=2/5$.

Таким образом, получена ранее вычисленная вероятность.

Теорема 2.1. *Вероятность появления одного из двух несовместных событий равна сумме вероятностей этих событий:*

$$P(A+B)=P(A)+P(B). \qquad (2.4)$$

Доказательство. Введем следующие обозначения: N — общее число возможных случаев в опыте; M_1 — число случаев, благоприятствующих наступлению события A; M_2 — число случаев, благоприятствующих наступлению события B.

Число случаев, благоприятствующих наступлению либо события A, либо события B, равно M_1+M_2. Так как

события A и B несовместны, то нет случаев, благоприятствующих наступлению событий A и B одновременно, следовательно,

$$P(A+B) = (M_1 + M_2)/N = M_1/N + M_2/N. \qquad (2.5)$$

Учитывая, что $P(A) = M_1/N$; $P(B) = M_2/N$, и подставляя эти выражения в равенство (2.5), получаем тождество.

Методом полной индукции можно обобщить теорему сложения на произвольное число попарно несовместных событий:

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i). \qquad (2.6)$$

Следствие 1. *Если несовместные события A_1, A_2, ..., A_n образуют полную группу, то сумма вероятностей этих событий равна единице, т. е.*

$$P(A_1) + P(A_2) + \cdots + P(A_n) = 1.$$

Доказательство. Используя свойство 3°, приведенное на с. 50, можно записать

$$P(A_1 + A_2 + \ldots + A_n) = 1.$$

Так как, по условию, события A_1, A_2,..., A_n — несовместные, то, применяя теорему сложения вероятностей к левой части равенства, получаем

$$P(A_1 + A_2 + \cdots + A_n) = P(A_1) + P(A_2) + \\ + \cdots + P(A_n) = 1. \quad \square$$

Следствие 2. *Сумма вероятностей противоположных событий равна единице:*

$$P(A) + P(\bar{A}) = 1.$$

Противоположные события образуют полную группу из двух несовместных событий. Поэтому это следствие можно вывести как частный случай из предыдущего следствия, не прибегая к отдельному доказательству. Последнее следствие часто используют при решении практических задач, когда оказывается легче вычислить вероятность противоположного события $\bar{A}$, чем вероятность прямого события A.

Если события совместны, то при сложении вероятностей этих событий следует учитывать вероятность их совместного появления. Докажем следующую теорему.

Теорема 2.2. Сложение вероятностей двух совместных событий. *Вероятность суммы двух совместных событий равна сумме вероятностей этих двух событий без вероятности их совместного появления:*

$$P(A+B)=P(A)+P(B)-P(AB). \qquad (2.7)$$

Доказательство. Используя геометрическую интерпретацию событий (см. рис. 2.1), можно события A, B и их сумму $A+B$ представить в виде сумм несовместных событий:

$$A=AB+A\bar{B},\ B=AB+\bar{A}B,$$

$$A+B=AB+\bar{A}B+A\bar{B}.$$

На основании теоремы сложения несовместных событий имеем следующие равенства:

$$P(A)=P(AB)+P(A\bar{B}), \qquad (2.8)$$

$$P(B)=P(AB)+P(\bar{A}B), \qquad (2.9)$$

$$P(A+B)=P(AB)+P(\bar{A}B)+P(A\bar{B}). \qquad (2.10)$$

Складывая левые и правые части равенств (2.8) и (2.9), получаем

$$P(A)+P(B)=2P(AB)+P(\bar{A}B)+P(A\bar{B}).$$

Отсюда

$$P(\bar{A}B)+P(A\bar{B})=P(A)+P(B)-2P(AB).$$

Подставляя левую часть этого равенства в равенство (2.10), получаем

$$P(A+B)=P(A)+P(B)-P(AB). \quad \square$$

Аналогично можно доказать, что вероятность суммы трех совместных событий

$$P(A+B+C)=P(A)+P(B)+P(C)-P(AB)-$$
$$-P(BC)-P(AC)+P(ABC). \qquad (2.11)$$

При большом числе событий вероятность появления хотя бы одного из них удобно вычислять по формуле

$$P(B)=1-P(\bar{B}),$$

где $B=A_1+A_2+\cdots+A_n$, $\bar{B}=A_1\bar{A}_2\cdots\bar{A}_n$.

§ 2.6. Теорема умножения вероятностей

Различают зависимые и независимые события. Два события называются *независимыми,* если появление одного из них не изменяет вероятности появления другого. Например, если в цехе работают две автоматические линии, по условиям производства не связанные между собой, то остановки этих линий — независимые события.

Несколько событий называют *независимыми в совокупности,* если любая их комбинация взаимно независима. События называют *зависимыми,* если появление одного из них изменяет вероятность появления другого. Пусть, например, две производственные установки связаны единым технологическим циклом. Тогда вероятность выхода из строя одной из них зависит от того, в каком состоянии находится другая. Вероятность одного события (B), вычисленная в предположении осуществления другого события (A), называется *условной вероятностью* события B и обозначается $P(B/A)$.

Условие независимости события B от события A записывают в виде $P(B/A)=P(B)$, а условие зависимости — в виде $P(B/A)\neq P(B)$. Рассмотрим пример вычисления условной вероятности события.

Пример 2.3. В ящике находятся пять сверл — два изношенных и три новых. Производят два последовательных извлечения сверл. Определить вероятность появления изношенного сверла при втором извлечении при условии, что извлеченное в первый раз сверло в ящик не возвращается.

Решение. Обозначим через A извлечение изношенного сверла в первом случае, а через $\bar{A}$ — извлечение нового. Тогда $P(A)=2/5$; $P(\bar{A})=3/5$.

Извлеченное сверло в ящик не возвращается, поэтому изменяется соотношение между изношенными и новыми сверлами, следовательно, вероятность извлечения изношенного сверла во втором случае зависит от того, какое событие осуществилось перед этим.

Пусть B — событие, означающее извлечение изношенного сверла во втором случае. Вероятности этого события могут быть такими: $P(B/A)=1/4$; $P(B/\bar{A})=2/4=1/2$.

Следовательно, вероятность события B зависит от того, произошло или нет событие A.

Теорема 2.3. *Вероятность произведения двух событий равна произведению вероятности одного из них на условную вероятность другого, вычисленную при условии, что первое событие произошло:*

$$P(AB)=P(A)\,P(B/A), \qquad (2.12)$$

или

$$P(AB)=P(B)\,P(A/B).$$

Доказательство. Рассмотрим опыт, сводящийся к схеме случаев. Пусть N — общее число возможных случаев в опыте. Предположим, что событию A благоприятны M случаев, а событию B — K случаев. Событие B наступает в предположении, что событие A уже произошло. Следовательно, совместному наступлению событий AB благоприятны из общего числа N случаев только K:

$$P(AB) = K/N.$$

При вычислении условной вероятности события B нужно учитывать, что общее возможное число случаев для этого события состоит только из тех случаев, которые благоприятствовали появлению события A:

$$P(B/A) = K/M, \quad P(A) = M/N.$$

Подставляя последние выражения в формулу (2.12), получаем тождество. □

Так, для рассмотренного выше примера вероятность двухкратного извлечения изношенных сверл $P(AB) = P(A)P(B/A) = 2/5 \cdot 1/4 = 1/10$.

Теорему 2.3 можно методом полной индукции распространить на произвольное число событий: *вероятность произведения нескольких событий равна произведению вероятности одного из этих событий на условные вероятности других. Условная вероятность каждого последующего события вычисляется в предположении, что все предыдущие события произошли:*

$$P(A_1 A_2 \cdots A_{n-1} A_n) = P(A_1) P(A_2/A_1) \cdots \times$$
$$\times P(A_n/A_1 A_2 \cdots A_{n-1}). \qquad (2.13)$$

Следствие 1. *Если появление события A не зависит от события B, то и появление события B не зависит от события A.*

Доказательство. Если появление события A не зависит от события B, то можно записать

$$P(A) = P(A/B). \qquad (2.14)$$

Используя две формы записи теоремы умножения, имеем $P(A)P(B/A) = P(B)P(A/B)$. Подставим в это тождество выражение (2.14):

$$P(A) P(B/A) = P(B) P(A).$$

Разделив обе части равенства на $P(A)$, получим

$$P(B/A) = P(B). \quad \square$$

Следствие 2. *Вероятность произведения двух независимых событий равна произведению вероятностей этих событий.*

Доказательство. Учитывая следствие 1, преобразуем формулу (2.12). В результате получим следующее равенство:

$$P(AB) = P(A)\,P(B). \quad \square$$

Это следствие может быть распространено на произвольное число событий, независимых в совокупности:

$$P(A_1 A_2 \cdots A_n) = P(A_1)\,P(A_2) \cdots P(A_n). \qquad (2.15)$$

Пример 2.4. На автоматическую линию сборки автомобильных колес поступают для запрессовки ободья и диски. Запрессовка доброкачественна, если детали по сопрягаемым диаметрам изготовлены без отклонения от допуска. Определить вероятность того, что запрессовка доброкачественна, если вероятность изготовления внутреннего диаметра обода в пределах допуска равна 0,98, а вероятность изготовления наружного диаметра диска в пределах допуска равна 0,95.

Решение. Пусть A — событие изготовления годного обода; B — событие изготовления годного диска. Технологические процессы изготовления ободьев и дисков независимы друг от друга, поэтому появление бракованного диска не влияет на изменение вероятности появления годного обода. Следовательно, искомая вероятность $P(AB) = P(A)P(B) = 0{,}98 \cdot 0{,}95 = 0{,}931$.

Пример 2.5. Известно, что 85% готовой продукции цеха шинпотреба является стандартной. Вероятность того, что она отличного качества, равна 0,51. Найти вероятность того, что наудачу взятое изделие окажется отличного качества.

Решение. Пусть A — событие, означающее, что взятое наудачу изделие стандартное; B — событие, означающее, что изделие отличного качества. Изделие может быть отличного качества, если оно стандартное. Поэтому из условия задачи следует, что $P(A) = 0{,}85$ и $P(B/A) = 0{,}51$. Тогда $P(AB) = P(A)P(B/A) = 0{,}85 \cdot 0{,}51 = 0{,}4335$.

§ 2.7. Формула полной вероятности

Теорема 2.4. *Если событие A может наступить только при условии появления одного из событий $B_1, B_2, \ldots, B_n$, образующих полную группу несовместных событий, то вероятность события A равна сумме произведений вероятностей каждого из событий $B_1, B_2, \ldots, B_n$ на соответствующую условную вероятность события A:*

$$P(A) = \sum_{i=1}^{n} P(B_i)\,P(A/B_i). \qquad (2.16)$$

Эту формулу называют *формулой полной вероятности.*

Доказательство. Формулу полной вероятности можно вывести на основании теоремы сложения и умножения вероятностей.

Согласно условию, событие A можно представить в виде несовместных комбинаций

$$A = B_1 A + B_2 A + \cdots + B_n A.$$

По теореме сложения вероятностей несовместных событий можно записать

$$P(A) = P(B_1 A) + P(B_2 A) + \cdots + P(B_n A).$$

По теореме умножения вероятностей имеем

$$P(A) = P(B_1) P(A/B_1) + P(B_2) P(A/B_2) + \\ + \cdots + P(B_n) P(A/B_n),$$

или

$$P(A) = \sum_{i=1}^{n} P(B_i) P(A/B_i). \quad \square$$

Пример 2.6. На сборочный конвейер поступают детали, изготовленные на трех станках. Производительность станков не одинакова. Первый дает 50% программы, второй — 30%, а третий — 20% Если в сборку попадает деталь, сделанная на первом станке, то вероятность получения годного узла равна 0,98. Для продукции второго и третьего станков соответствующие вероятности равны 0,95 и 0,8. Определить вероятность того, что узел, сходящий с конвейера, годный.

Решение. Пусть A — событие, означающее годность собранного узла; B_1, B_2 и B_3 — события, означающие, что детали сделаны соответственно на первом, втором и третьем станках. Тогда имеем: $P(B_1) = 0{,}5$; $P(B_2) = 0{,}3$; $P(B_3) = 0{,}2$; $P(A/B_1) = 0{,}98$; $P(A/B_2) = 0{,}95$; $P(A/B_3) = 0{,}8$. Искомая вероятность

$$P(A) = P(B_1) P(A/B_1) + P(B_2) P(A/B_2) + P(B_3) P(A/B_3) = \\ = 0{,}5 \cdot 0{,}98 + 0{,}3 \cdot 0{,}95 + 0{,}2 \cdot 0{,}8 = 0{,}935.$$

§ 2.8. Формулы Байеса или теорема гипотез

Формулы Байеса применяют при решении практических задач, когда событие A, появляющееся совместно с каким-либо из событий $B_1, B_2, \ldots, B_n$, которые образуют полную группу несовместных событий (эти события иногда называются гипотезами), произошло и требуется произвести количественную переоценку вероятностей событий $B_1, B_2, \ldots, B_n$. Априорные (до опыта) вероятности $P(B_1), P(B_2), \ldots, P(B_n)$ известны. Требуется вычислить

апостериорные (после опыта) вероятности, т. е., по существу, нужно найти условные вероятности $P(B_1/A)$, $P(B_2/A), ..., P(B_n/A)$.

Пусть событие A может наступить при условии появления одного из несовместных событий B_1, B_2, ..., B_j, ..., B_n, образующих полную группу. Вероятность совместного наступления события A с любым из этих событий (пусть это будет событие B_j) согласно теореме умножения вероятностей определится следующим образом:

$$P(AB_j) = P(A)\,P(B_j/A),\ P(AB_j) = P(B_j)\,P(A/B_j).$$

Левые части этих выражений равны, следовательно, равны и правые части, поэтому

$$P(A)\,P(B_j/A) = P(B_j)\,P(A/B_j).$$

Решая это уравнение относительно $P(B_j/A)$ при условии, что $P(A) \neq 0$, получаем

$$P(B_j/A) = \frac{P(B_j)\,P(A/B_j)}{P(A)}.$$

Раскрывая в последнем равенстве $P(A)$ по формуле полной вероятности, имеем

$$P(B_j/A) = \frac{P(B_j)\,P(A/B_j)}{\sum_{i=1}^{n} P(B_i)\,P(A/B_i)}. \qquad (2.17)$$

Эта формула носит название *формулы Байеса.*

Пример 2.7. Пользуясь данными примера 2.6, найти вероятность того, что в сборку попала деталь, изготовленная на первом, втором и третьем станках соответственно, если узел, сходящий с конвейера, годный.

Р е ш е н и е. Расчет условных вероятностей произведем по формулам Байеса. Имеем: для первого станка

$$P(B_1/A) = \frac{P(B_1)\,P(A/B_1)}{P(A)} = \frac{0{,}5 \cdot 0{,}98}{0{,}935} = 0{,}525;$$

для второго станка

$$P(B_2/A) = \frac{P(B_2)\,P(A/B_2)}{P(A)} = \frac{0{,}3 \cdot 0{,}95}{0{,}935} = 0{,}304;$$

для третьего станка

$$P(B_3/A) = \frac{P(B_3)\,P(A/B_3)}{P(A)} = \frac{0{,}2 \cdot 0{,}8}{0{,}935} = 0{,}171.$$

§ 2.9. Схема испытаний Бернулли

На практике приходится сталкиваться с задачами, которые можно представить в виде многократно повторяющихся испытаний, в результате каждого из которых может появиться или не появиться событие A. При этом представляет интерес исход не каждого отдельного испытания, а общее число появлений события A в результате определенного количества испытаний. В подобных задачах нужно уметь определять вероятность любого числа m появлений события A в результате n испытаний. Рассмотрим случай, когда испытания являются независимыми и вероятность появления события A в каждом испытании одинакова.

Если вероятность события A в каждом испытании не зависит от исходов других испытаний, то такие испытания называют *независимыми относительно события A*. Если независимые испытания производятся в одинаковых условиях, то вероятность события A в этих испытаниях одна и та же. При изменении условий испытаний вероятность события A от испытания к испытанию меняется. Примером независимых испытаний может служить проверка на годность изделий, взятых по одному из ряда партий. Если в этих партиях процент брака одинаковый, то вероятность того, что отобранное изделие будет бракованным, в каждом случае является постоянным числом.

Вывод формулы Бернулли. Пусть производят n независимых испытаний, в каждом из которых событие A может либо появиться с вероятностью p, либо не появиться с вероятностью $q=1-p$. Рассмотрим событие B_m, состоящее в том, что событие A в этих n испытаниях наступит ровно m раз и, следовательно, не наступит ровно $n-m$ раз. Обозначим через A_i $(i=1, 2,..., n)$ появление события A, через $\bar{A}_i$ — непоявление события A в i-м испытании. В силу постоянства условий испытаний имеем

$$P(A_1) = P(A_2) = \cdots = P(A_n) = p;$$

$$P(\bar{A}_1) = P(\bar{A}_2) = \cdots = P(\bar{A}_n) = 1 - p = q.$$

Событие A может появиться m раз в разных последовательностях или комбинациях, чередуясь с противоположным событием $\bar{A}$. Число возможных комбинаций такого рода равно числу сочетаний из n элементов по m,

т. е. C_n^m. Следовательно, событие B_m можно представить в виде суммы различных комбинаций событий, несовместных между собой, причем число слагаемых равно C_n^m:

$$B_m = A_1 A_2 \cdots A_m \bar{A}_{m+1} \cdots \bar{A}_n + \bar{A}_1 A_2 A_3 \cdots \bar{A}_{n-1} A_n + \\ + \bar{A}_1 \bar{A}_2 \cdots \bar{A}_{n-m} A_{n-m+1} \cdots A_n, \qquad (2.18)$$

где в каждое произведение событие A входит m раз, а $\bar{A}$ входит $n-m$ раз. Вероятность каждой последовательности событий, входящей в формулу (2.18), по теореме умножения вероятностей для независимых событий, равна $p^m q^{n-m}$. Так как общее число таких последовательностей равно C_n^m, то, используя теорему сложения вероятностей для несовместных событий, получим вероятность события B_m (обозначим ее $P_{m,n}$): $P_{m,n} = C_n^m p^m q^{n-m}$, или

$$P_{m,n} = \frac{n!}{m!\,(n-m)!}\, p^m\, q^{n-m}. \qquad (2.19)$$

Полученную формулу называют *формулой Бернулли*, а повторяющиеся испытания, удовлетворяющие условию независимости и постоянства вероятностей появления в каждом из них события A, называют *испытаниями Бернулли* или *схемой Бернулли*.

События B_m ($m=0, 1, \ldots, n$) состоящие в различном числе появлений события A в n испытаниях, несовместны и образуют полную группу. Следовательно,

$$\sum_{m=0}^{n} P_{m,n} = \sum_{m=0}^{n} C_n^m\, p^m\, q^{n-m} =$$

$$= q^n + C_n^1 p q^{n-1} + \cdots + C_n^m p^m\, q^{n-m} + \cdots + p^n = 1. \qquad (2.20)$$

Нетрудно заметить, что правая часть последнего равенства представляет собой члены разложения бинома

$$(q+p)^n = q^n + C_n^1 p q^{n-1} + \cdots + C_n^m\, p^m\, q^{n-m} + \cdots + p^n = 1.$$

Пример 2.8. Вероятность выхода за границы поля допуска при обработке плунжера на токарном станке равна 0,07. Определить вероятность того, что из пяти наудачу отобранных в течение смены деталей у одной размеры диаметра не соответствуют заданному допуску.

Р е ш е н и е. Условие задачи удовлетворяет требованиям схемы Бернулли. Поэтому по формуле (2.19), полагая $n=5$, $m=1$, $p=0{,}07$, получаем $P_{1,5} = C_5^1 (0{,}07)^1 (0{,}93)^4 = 0{,}2618$.

§ 2.10. Вероятнейшее число появлений события

Вероятнейшим числом появления события A в n независимых испытаниях называется такое число m_0, для которого вероятность, соответствующая этому числу, не меньше вероятности каждого из остальных возможных чисел появления события A. Для определения вероятнейшего числа не обязательно вычислять вероятности всех возможных чисел появлений события, достаточно знать число испытаний n и вероятность появления события A в отдельном испытании. Обозначим вероятность, соответствующую вероятнейшему числу m_0, через $P_{m_0,n}$. Тогда согласно определению вероятнейшего числа имеем $P_{m_0,n} \geqslant P_{m_0+1,n}$, $P_{m_0,n} \geqslant P_{m_0-1,n}$.

Решая эти неравенства относительно m_0 и учитывая при этом формулу (2.19), получаем двойное неравенство, которое используется для определения вероятнейшего числа:

$$np - q \leqslant m_0 \leqslant np + p. \qquad (2.21)$$

Так как длина интервала, определяемого неравенством (2.21), равна единице: $(np+p)-(np-q)=p+q=1$, а событие может произойти в n испытаниях только целое число раз, то следует иметь в виду, что:

1) если $np-q$ — целое число, то существуют два значения вероятнейшего числа, а именно: $m_0=np-q$ и $m_0'=np-q+1=np+p$;

2) если $np-q$ — дробное число, то существует одно вероятнейшее число, а именно единственное целое, заключенное между дробными числами, полученными из неравенства (2.21);

3) если np — целое число, то существует одно вероятнейшее число, а именно $m_0=np$.

При больших значениях n пользоваться биномиальной формулой (2.19) для подсчета вероятности, соответствующей вероятнейшему числу, неудобно. Если в равенство (2.19) подставить формулу Муавра—Стирлинга

$$n! \approx n^n e^{-n} \sqrt{2\pi n},$$

справедливую для достаточно больших n, и принять вероятнейшее число равным $m_0=np$, то получим формулу для приближенного вычисления вероятности, соответствующей вероятнейшему числу:

$$P_{m_0,n} \approx \frac{n^n e^{-n}\sqrt{2\pi n}\, p^{np} q^{nq}}{(np)^{np} e^{-np}\sqrt{2\pi np}\,(nq)^{nq} e^{-nq}\sqrt{2\pi nq}} = \frac{1}{\sqrt{2\pi npq}}. \quad (2.22)$$

С увеличением n возрастает точность этой формулы.

Пример 2.9. Известно, что $^1/_{15}$ продукции, поставляемой заводом на торговую базу, не удовлетворяет всем требованиям стандарта. На базу завезена партия изделий, объем которой составляет 250 шт. Найти вероятнейшее число изделий, удовлетворяющих требованиям стандарта, и вычислить вероятность того, что в этой партии окажется вероятнейшее число изделий.

Решение. По условию, $n=250$; $q=1/15$, $p=1-1/15=14/15$. Согласно неравенству (2.21), имеем $250\cdot 14/15-1/15<m_0<250\times$ $\times 14/15+14/15$, откуда $233{,}26<m_0<234{,}26$.

Следовательно, вероятнейшее число изделий, удовлетворяющих требованиям стандарта, в партии из 250 шт. равно 234. Подставляя данные в формулу (2.22), найдем вероятность наличия в партии вероятнейшего числа изделий:

$$P_{234,250} \approx 1/\sqrt{2\pi\cdot 250\cdot 14/15\cdot 1/15} = 0{,}1012.$$

Глава 3

Распределение случайных величин

§ 3.1. Случайная величина. Задание закона ее распределения

Случайной называют величину, которая принимает в результате испытания то или иное (но при этом только одно) возможное значение, заранее неизвестное, меняющееся от испытания к испытанию и зависящее от случайных обстоятельств. В отличие от случайного события, являющегося качественной характеристикой случайного результата испытания, случайная величина характеризует результат испытания количественно. Примерами случайной величины могут служить размер обрабатываемой детали, содержание химических элементов в чугуне и т. д. Случайные величины могут быть дискретными и непрерывными.

Дискретной называют такую случайную величину, которая принимает конечное или бесконечное счетное множество значений. Это, например, число дефектных изделий в партии; число вызовов, поступающих на телефонную станцию в течение суток; число отказов элементов устройства за определенный промежуток времени при испытании его на надежность и т. д.

Непрерывной называют такую случайную величину, которая может принимать любые значения из некоторого конечного или бесконечного интервала. Очевидно, число возможных значений непрерывной случайной величины бесконечно. Время безотказной работы как отдельных элементов системы, так и всей системы в целом, погрешность измерения физических величин, погрешность изготовления деталей на станке — примеры непрерывных случайных величин.

Случайные величины обычно обозначают заглавными

буквами конца латинского алфавита — X, Y, Z,..., а их возможные значения — соответствующими малыми буквами — x, y, z,...

Для задания случайной величины недостаточно перечислить все ее возможные значения. Необходимо также знать, как часто могут появиться те или иные значения в результате испытаний при одних и тех же условиях, т. е. нужно задать вероятности их появления.

Пусть X — дискретная случайная величина, принимающая в результате испытания следующие возможные значения:

$$X = x_1;\ X = x_2;\ \ldots;\ X = x_n. \qquad (3.1)$$

Перечисленные значения случайной величины можно рассматривать как события, появляющиеся при испытаниях. Обозначим вероятности этих событий через p с соответствующими индексами: $P(X=x_1)=p_1$; $P(X=x_2)=p_2$;...; $P(X=x_n)=p_n$. События (3.1) образуют полную группу n несовместных событий, следовательно, сумма вероятностей всех возможных значений случайной величины X равна единице:

$$\sum_{i=1}^{n} P(X = x_i) = \sum_{i=1}^{n} p_i = 1.$$

Распределением (законом) случайной величины называется всякое соотношение между возможными значениями случайной величины и соответствующими им вероятностями. Про случайную величину говорят, что она *подчиняется* данному закону распределения. Две случайные величины называют *независимыми*, если распределение вероятностей одной из них не зависит от того, какие возможные значения приняла другая величина. В противном случае случайные величины называют *зависимыми*.

Несколько случайных величин называют *взаимно независимыми*, если законы распределения любого числа из них не зависят от того, какие возможные значения приняли остальные величины. Распределение случайной величины может быть задано в виде таблицы, в виде функции распределения и в виде плотности распределения. Таблица, содержащая возможные значения случайной величины и соответствующие вероятности, является простейшей формой задания распределения случайной величины:

X	x_1	x_2	x_3	...	x_{n-1}	x_n
P	p_1	p_2	p_3	...	p_{n-1}	p_n

(3. 2)

Табличное задание распределения может быть использовано только для дискретной случайной величины с конечным числом возможных значений. Непрерывная случайная величина имеет бесчисленное множество возможных значений, поэтому перечислить их в таблице невозможно. Кроме того, как будет показано далее, каждое отдельное значение непрерывной случайной величины обладает нулевой вероятностью. Табличную форму задания распределения случайной величины называют также *рядом распределения.*

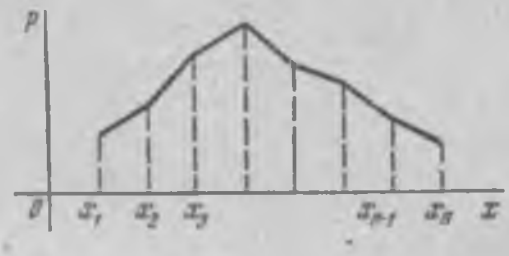

Рис. 3.1

При графическом изображении ряда распределения в прямоугольной системе координат по оси абсцисс откладывают все возможные значения случайной величины, а по оси ординат—соответствующие вероятности. Затем строят точки $(x_i; p_i)$ и соединяют их прямолинейными отрезками (рис. 3.1). Полученную фигуру называют *многоугольником распределения.*

Функция распределения является наиболее общей формой задания закона случайной величины. Она используется для задания как дискретных, так и непрерывных случайных величин. Обычно ее обозначают $F(x)$. *Функция распределения* определяет вероятность того, что случайная величина X примет значение, меньшее фиксированного действительного числа x, т. е.

$$F(x) = P(X < x). \qquad (3.3)$$

Вероятность того, что $X < x$, зависит от x, следовательно, $F(x)$ является функцией от x. Поэтому $F(x)$ и называется *функцией* распределения. В литературе встречаются также термины: «интегральная функция распределения» и «интегральный закон распределения».

Геометрическая интерпретация функции распределения очень проста. Если случайную величину рассматривать как случайную точку X оси Ox (рис. 3.2), которая в результате испытания может занять то или иное положение на этой оси, то функция распределения $F(x)$ есть

Рис. 3.2

вероятность того, что случайная точка X в результате испытания попадает левее точки x. Для дискретной случайной величины X, которая может принимать значения $x_1, x_2, \ldots, x_n$, функция распределения имеет вид

$$F(x) = \sum_{x_i < x} P(X = x_i), \tag{3.4}$$

где неравенство $x_i < x$ под знаком суммы означает, что суммирование распространяется на все те значения x_i, которые по своей величине меньше x.

Построим функцию распределения случайной величины X, ряд распределения которой имеет вид (3.2): при $x \leqslant x_1$

$$F(x) = P(X < x_1) = 0;$$

при $x_1 < x \leqslant x_2$

$$F(x) = P(X < x_2) = P(X = x_1) = p_1;$$

при $x_2 < x \leqslant x_3$

$$F(x) = P(X < x_3) = P(X = x_1) + P(X = x_2) = p_1 + p_2;$$

при $x_3 < x \leqslant x_4$

$$F(x) = P(X < x_4) = P(X = x_1) + P(X = x_2) + P(X = x_3) = p_1 + p_2 + p_3;$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

при $x_{n-1} < x \leqslant x_n$

$$F(x) = P(X < x_n) = P(X = x_1) + P(X = x_2) + P(X = x_3) + \cdots + P(X = x_{n-1}) = p_1 + p_2 + p_3 + \cdots + p_{n-1};$$

при $x > x_n$

$$F(x) = P(X = x_1) + P(X = x_2) +$$
$$+ P(X = x_3) + \cdots + P(X = x_{n-1}) + P(X = x_n) =$$
$$= p_1 + p_2 + p_3 + \cdots + p_{n-1} + p_n = 1.$$

Функция распределения имеет скачок в точках, где случайная величина принимает конкретные значения, указанные в ряду распределения. В интервалах между значениями случайной величины функция $F(x)$ постоянна. Величина скачка в точке x_i равна вероятности p_i. Сумма всех скачков функции распределения равна единице. График функции распределения дискретной случайной величины — разрывная ступенчатая ломаная линия (рис. 3.3).

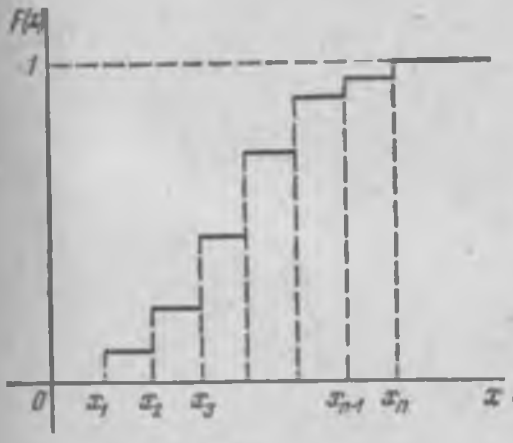

Рис. 3.3

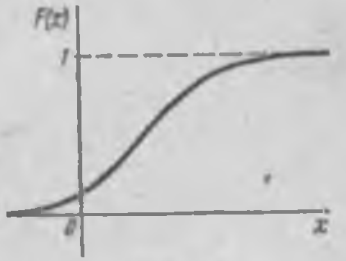

Рис. 3.4

Непрерывная случайная величина имеет непрерывную функцию распределения; график этой функции имеет форму плавной кривой (рис. 3.4). Рассмотрим общие свойства функции распределения.

1°. *Функция распределения $F(x)$ есть неотрицательная функция, заключенная между нулем и единицей:*

$$0 \leqslant F(x) \leqslant 1.$$

Это свойство вытекает из определения функции распределения как вероятности осуществления события, заключающегося в выполнении неравенства $X < x$. Функция распределения, как и всякая вероятность, есть величина безразмерная.

2°. *Функция распределения случайной величины есть неубывающая функция, т. е. при* $x_2 \geqslant x_1$

$$F(x_2) \geqslant F(x_1),$$

так как вероятность попасть на отрезок $(-\infty; x_2)$ не меньше, чем вероятность попасть на отрезок $(-\infty, x_1)$ при $x_2 \geqslant x_1$.

3°. *На минус-бесконечности функция распределения* $F(x)$ *равна нулю, а на плюс-бесконечности функция распределения равна единице, т. е.*

$$F(-\infty) = 0;\ F(+\infty) = 1.$$

Свойство становится очевидным при геометрической интерпретации функции распределения (см. рис. 3.2). Если точка x неограниченно перемещается влево, то попадание случайной точки X левее x в пределе становится невозможным событием. Поэтому естественно полагать, что вероятность этого события стремится к нулю, т. е. можно записать $F(-\infty)=0$.

При неограниченном перемещении точки x вправо попадание случайной точки X левее x в пределе — достоверное событие. Поэтому естественно полагать, что вероятность этого события стремится к единице, т. е. $F(+\infty)=1$.

Кратко свойство функции распределения можно сформулировать так: *функция распределения является неотрицательной неубывающей функцией, удовлетворяющей условиям* $F(-\infty)=0$ *и* $F(+\infty)=1$. Возможно и обратное утверждение: *функция, удовлетворяющая перечисленным условиям, может рассматриваться как функция распределения некоторой случайной величины.*

Следует заметить, что, в то время как каждая случайная величина однозначно определяет функцию распределения, одну и ту же функцию распределения могут иметь различные случайные величины. Можно привести более корректное определение случайной величины, а именно: *случайной величиной называется переменная величина, значения которой зависят от случая и для которой определена функция распределения вероятностей.*

При решении ряда задач теории вероятностей важнее знать не вероятность $P(X=x)$, а вероятность попадания случайной величины в заданный интервал.

Вероятность попадания случайной величины в интервал (α, β) *равна разности значений функции распределения на концах этого интервала:*

$$P(\alpha \leqslant X < \beta) = F(\beta) - F(\alpha). \qquad (3.5)$$

Доказательство. Рассмотрим следующие три события: событие A, состоящее в том, что $X < \beta$; событие B, состоящее в том, что $X < \alpha$; событие C, состоящее в том, что $\alpha \leqslant X < \beta$.

Можно записать, что

$$P(A) = P(X < \beta) = F(\beta);$$

$$P(B) = P(X < \alpha) = F(\alpha), \quad P(C) = P(\alpha \leqslant X < \beta)^*.$$

Очевидно, что событие A представляет собой сумму двух несовместных событий B и C: $A = B + C$. По теореме сложения вероятностей,

$$P(A) = P(B) + P(C),$$

или

$$F(\beta) = F(\alpha) + P(\alpha \leqslant X < \beta), \qquad (3.6)$$

откуда

$$P(\alpha \leqslant X < \beta) = F(\beta) - F(\alpha). \quad \square$$

Следствие. *Вероятность любого отдельного значения непрерывной случайной величины равна нулю.*

Полагая, что $\beta \to \alpha$, будем неограниченно уменьшать интервал $[\alpha, \beta)$. В пределе вместо вероятности попадания случайной величины X в интервал $[\alpha, \beta]$ получим вероятность того, что эта величина принимает отдельно взятое значение α:

$$P(X = \alpha) = \lim_{\beta \to \alpha} P(\alpha \leqslant X < \beta) = \lim_{\beta \to \alpha} [F(\beta) - F(\alpha)]. \qquad (3.7)$$

Значение этого предела зависит от того, является ли функция $F(x)$ в точке α непрерывной или же терпит разрыв. Если в точке α функция $F(x)$ имеет разрыв, то предел (3.7) равен значению скачка функции $F(x)$ в точке α. Если же функция $F(x)$ в точке α непрерывна, то этот предел равен нулю.

Так как непрерывная случайная величина X имеет непрерывную функцию распределения $F(x)$, то из равенства (3.7) следует, что вероятность любого отдельного значения непрерывной случайной величины равна нулю. На первый взгляд, этот вывод кажется парадоксальным, так как события, вероятности которых равны нулю, в пре-

* Условимся левый конец включать, а правый — не включать в интервал $[\alpha, \beta)$.

дыдущей главе рассматривались нами как невозможные. Из изложенного в настоящем параграфе следует, что нулевой вероятностью обладают не только невозможные, но и возможные события. Действительно, событие, состоящее в том, что непрерывная случайная величина X принимает значение α, возможно, но вероятность этого события равна нулю. Этим непрерывная случайная величина отличается от дискретной (так называемый парадокс непрерывности).

На основании этого следствия для непрерывной случайной величины можно записать формулу (3.5), не включая в рассматриваемый интервал $(\alpha, \beta]$ левый его конец:

$$P(\alpha < X < \beta) = F(\beta) - F(\alpha). \qquad (3.8)$$

Непрерывную случайную величину можно задать не только интегральной функцией распределения, но и дифференциальной функцией. Рассмотрим эту форму задания распределения случайной величины. Пусть имеется непрерывная случайная величина X с функцией распределения $F(x)$. Вероятность попадания этой случайной величины на элементарный участок $[x; x+\Delta x]$ через интегральную функцию определится, согласно формуле (3.5), так:

$$P(x < X + \Delta x) = F(x + \Delta x) - F(x).$$

Разделим обе части этого равенства на Δx:

$$\frac{P(x < X < x + \Delta x)}{\Delta x} = \frac{F(x + \Delta x) - F(x)}{\Delta x}.$$

Полученное отношение называют *средней вероятностью*, приходящейся на единицу длины этого участка. Считая функцию распределения $F(x)$ дифференцируемой, перейдем в последнем равенстве к пределу при $\Delta x \to 0$. В результате получим производную от функции распределения, которую и называют дифференциальной функцией распределения:

$$\lim_{\Delta x \to 0} \frac{P(x < X < x + \Delta x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = F'(x). \qquad (3.9)$$

Введем обозначение: $f(x) = F'(x)$. Итак, *дифференциальной функцией распределения* $f(x)$ является первая производная от интегральной функции. Иногда функцию $f(x)$ называют *дифференциальным законом* распределения случайной величины X, или *плотностью*, или *функцией плотности* распределения вероятности. Заметим, что для описания распределения вероятностей дискретной слу-

чайной величины дифференциальная функция неприменима. Кривая, изображающая дифференциальную функцию распределения $f(x)$ случайной величины, называется *кривой распределения* (рис. 3.5). Выделим на оси абсцисс элементарный участок dx, примыкающий к точке x (рис. 3.6). Вероятность попадания случайной величины на этот элементарный участок с точностью до бесконечно малых высшего порядка равна $f(x)dx$. Геометрически это площадь элементарного прямоугольника с высотой $f(x)$ и основанием dx. Величина $f(x)dx$ называется *элементом вероятности*. Теперь можно дать более строгое определение непрерывной случайной величины: *случайная величина X называется непрерывной, если функция распределения $F(x)$ непрерывна на всей оси Ox, а плотность распределения $f(x)$ существует везде, за исключением, быть может, конечного числа точек.*

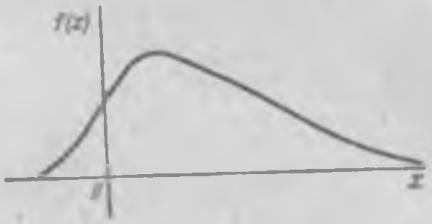

Рис. 3.5

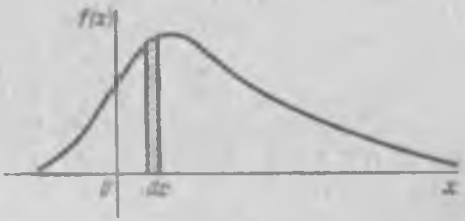

Рис. 3.6

Следует иметь в виду, что функция распределения $F(x)$, как всякая вероятность, есть безразмерная величина, а размерность плотности распределения $f(x)$ обратна размерности случайной величины.

Вероятность попадания непрерывной случайной величины X в интервал $[\alpha, \beta]$ равна определенному интегралу от дифференциальной функции, взятому в пределах от α до β:

$$P(\alpha < X < \beta) = \int_{\alpha}^{\beta} f(x)\,dx. \qquad (3.10)$$

Доказательство. На основании (3.8), учитывая следствие, сформулированное для непрерывной случайной величины, имеем

$$P(\alpha<X<\beta)=F(\beta)-F(\alpha).$$

Согласно основной теореме интегрального исчисления можно записать

$$\int_{\alpha}^{\beta} f(x)\,dx=F(\beta)-F(\alpha).$$

Таким образом,

$$P(\alpha<X<\beta)=\int_{\alpha}^{\beta} f(x)\,dx.$$

Геометрически это означает, что вероятность попадания непрерывной случайной величины X в интервал $[\alpha, \beta]$ численно равна площади криволинейной трапеции, заштрихованной на рис. 3.7.

Интегральная функция распределения может быть выражена через дифференциальную по формуле

$$F(x)=\int_{-\infty}^{x} f(x)\,dx. \qquad (3.11)$$

Доказательство. По определению интегральной функции,

$$F(x)=P(X<x)=P(-\infty<X<x). \qquad (3.12)$$

Полагая в формуле (3.10) $\alpha=-\infty$, $\beta=x$, имеем

$$P(-\infty<X<x)=\int_{-\infty}^{x} f(x)\,dx.$$

Заменив в последнем выражении $P(-\infty<X<x)$ на $F(x)$, в силу формулы (3.12) получим

$$F(x)=\int_{-\infty}^{x} f(x)\,dx.$$

На графике плотности распределения функции распределения соответствует площадь, заштрихованная на рис. 3.8.

Рассмотрим свойства дифференциальной функции распределения.

1°. *Дифференциальная функция распределения неотрицательна:*

$$f(x)\geqslant 0.$$

Это свойство непосредственно вытекает из определения дифференциальной функции как производной от неубывающей функции распределения $F(x)$. Геометрически оно означает, что кривая распределения расположена либо над осью абсцисс, либо на этой оси.

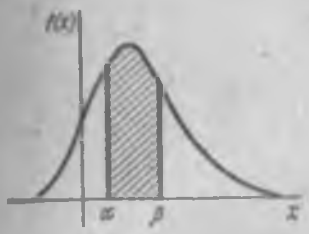

Рис. 3.7

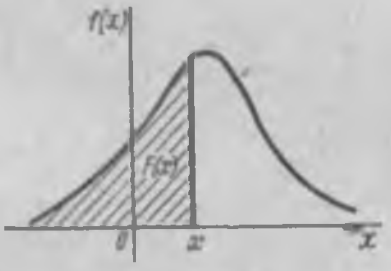

Рис. 3.8

2°. *Интеграл в бесконечных пределах от дифференциальной функции равен единице:*

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Доказательство. Полагая в равенстве (3.11) $x=\infty$ и учитывая, что $F(+\infty)=1$, получаем

$$F(+\infty) = \int_{-\infty}^{\infty} f(x)\,dx = 1. \quad \square$$

Геометрически это означает, что площадь, ограниченная кривой распределения и осью абсцисс, равна единице. Если случайная величина X распределена в более узких границах, чем $(-\infty;\ +\infty)$, например, $(a,\ b)$, то

$$\int_{a}^{b} f(x)\,dx = 1.$$

Пример 3.1. Вероятность выхода за границу поля допуска при шлифовании торцов плашек равна 0,25. Для контроля стабильности процесса через равные промежутки времени отбираются пять плашек. Определить вероятность того, что в выборке окажется меньше двух плашек, размеры которых выходят за границу поля допуска. Вычислить функцию распределения числа плашек, размеры которых выходят за границу поля допуска. Построить многоугольник распределения и график интегральной функции.

Решение. Обозначим число плашек, размеры которых выходят за границу поля допуска, через X, тогда возможные значения

случайной величины X таковы: $x_1=0$, $x_2=1$, $x_3=2$, $x_4=3$, $x_5=4$, $x_6=5$. Вероятность возможных значений случайной величны определим по формуле Бернулли:

$$P_{n,m} = C_n^m p^m q^{n-m}.$$

В нашем случае $n=5$; $m=x_i$, $i=1, 2, ..., 6$; $p=0{,}25$; $q=0{,}75$. Составим ряд распределения:

X	0	1	2	3	4	5
P	0,2373	0,3955	0,2637	0,0389	0,0146	0,0010

Многоугольник распределения изображен на рис. 3.9.

Построим функцию распределения случайной величины X. Воспользуемся формулой (3.4): при $x<0$

$$F(x) = P(X < 0) = 0;$$

при $0<x\leqslant 1$

$$F(x) = \sum_{x_i<1} P(X = x_i) = P(X = 0) = 0{,}2373;$$

при $1<x\leqslant 2$

$$F(x) = \sum_{x_i<2} P(X = x_i) = P(X = 0) + P(X = 1) = 0{,}2373 + 0{,}3955 = 0{,}6328;$$

при $2<x\leqslant 3$

$$F(x) = \sum_{x_i<3} P(X = x_i) = P(X = 0) + P(X = 1) + P(X = 2) = 0{,}2373 + 0{,}3955 + 0{,}2637 = 0{,}8965;$$

при $3<x\leqslant 4$

$$F(x) = \sum_{x_i<4} P(X = x_i) = P(X = 0) + P(X=1) + P(X=2) + P(X=3) = 0{,}2373 + 0{,}3955 + 0{,}2637 + 0{,}0879 = 0{,}9844;$$

при $4<x\leqslant 5$

$$F(x) = \sum_{x_i<5} P(X = x_i) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 0{,}9990;$$

при $x>5$

$$F(x) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) = 1.$$

Событие, состоящее в том, что в выборке должно быть меньше двух плашек, размеры которых выходят за границу поля допуска, означает, что случайная величина X должна удовлетворять неравенству $0<X<2$. Имеем

$$P(0 < X < 2) = F(2) = P(X = 0) + P(X = 1) = 0{,}2373 + 0{,}3955 = 0{,}6328.$$

График интегральной функции распределения представлен на рис. 3.10.

Пример 3.2. Случайная величина X подчинена закону распределения с плотностью

$$f(x)=\begin{cases} 0 & \text{при } x<0,\\ a\sin x & \text{при } 0<x<\pi,\\ 0 & \text{при } x>\pi. \end{cases}$$

Определить коэффициент a; построить график плотности распределения. Найти вероятность попадания случайной величины на участок от 0 до $\pi/2$. Определить интегральную функцию и построить ее график.

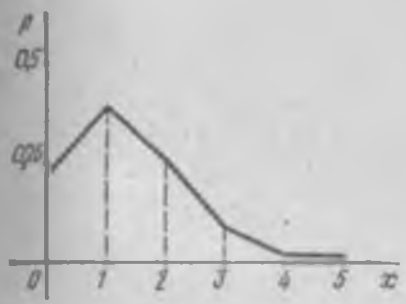

Рис. 3.9

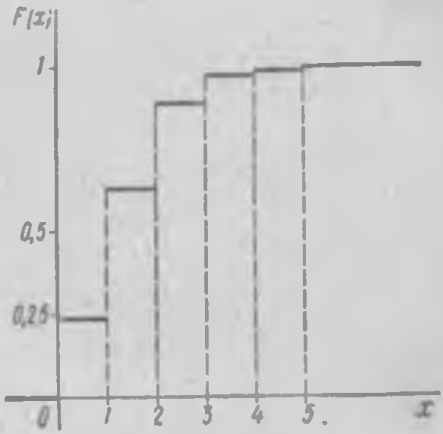

Рис. 3.10

Решение. Площадь, ограниченная кривой распределения, численно равна

$$\int_{-\infty}^{\infty} f(x)\,dx=\int_{0}^{\pi} a\sin x\,dx=-a\cos x\Big|_{0}^{\pi}=2a.$$

Учитывая свойство 2° плотности распределения, находим $a=1/2$. Следовательно, плотность распределения может быть выражена так:

$$f(x)=\begin{cases} (\sin x)/2 & \text{при } 0<x<\pi,\\ 0 & \text{при } x<0 \text{ и } x>\pi. \end{cases}$$

График плотности распределения изображен на рис 3.11. По формуле (3.10) имеем

$$P(0<X<\pi/2)=\int_{0}^{\pi/2} {}^{1}/_{2}\sin x\,dx=-{}^{1}/_{2}\cos x\Big|_{0}^{\pi/2}=$$

$$=-(\cos\pi/2-\cos 0)/2=1/2.$$

Для определения интегральной функции воспользуемся формулой (3.11)

$$F(x) = \int_0^x {}^1/_2 \sin x\,dx = -{}^1/_2 \cos x \Big|_0^x = {}^1/_2 - {}^1/_2 \cos x.$$

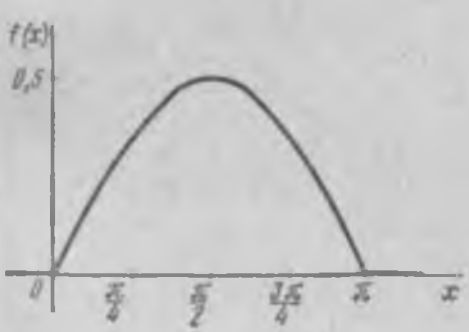

Рис. 3.11

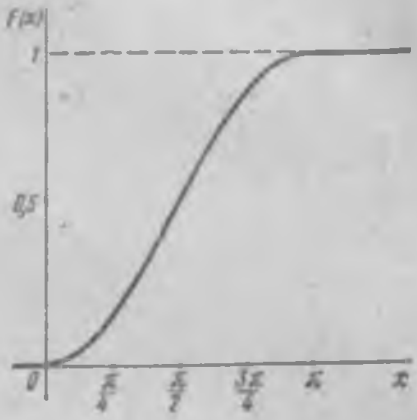

Рис. 3.12

Таким образом,

$$F(x) = \begin{cases} 0 & \text{при } x < 0, \\ {}^1/_2 - {}^1/_2 \cos x & \text{при } 0 < x < \pi, \\ 1 & \text{при } x > \pi. \end{cases}$$

График интегральной функции изображен на рис. 3.12.

§ 3.2. Числовые характеристики случайной величины

Распределение вероятностей полностью характеризует случайную величину. Но при решении ряда практических задач нет необходимости знать все возможные значения случайной величины и соответствующие им вероятности, а удобнее пользоваться некоторыми количественными показателями, которые давали бы в сжатой форме достаточную информацию о случайной величине. Такие показатели называют *числовыми характеристиками случайной величины*. Основными из них являются математическое ожидание, дисперсия, моменты различных порядков, мода и медиана.

Математическое ожидание характеризует положение случайной величины на числовой оси, определяя собой

центр распределения (некоторое среднее значение, около которого сосредоточены все возможные значения случайной величины). Поэтому математическое ожидание иногда называют просто средним значением случайной величины. Рассмотрим дискретную случайную величину X, принимающую значения $x_1, x_2, \ldots, x_n$ с вероятностями $p_1, p_2, \ldots, p_n$. Определим среднюю арифметическую значений случайной величины, взвешенных по вероятностям их появлений. Таким образом вычислим среднее значение случайной величины или ее математическое ожидание, которое будем обозначать $M(x)$ или $M[X]$:

$$M(x) = \frac{x_1 p_1 + x_2 p_2 + \cdots + x_n p_n}{p_1 + p_2 + \cdots + p_n} = \sum_{i=1}^{n} x_i p_i \Big/ \sum_{i=1}^{n} p_i.$$

Учитывая, что $\sum_{i=1}^{n} p_i = 1$, получаем

$$M(x) = \sum_{i=1}^{n} x_i p_i. \tag{3.13}$$

Если дискретная случайная величина X принимает бесконечное счетное множество значений, то ее математическое ожидание

$$M(x) = \sum_{i=1}^{\infty} x_i p_i.$$

Итак, *математическим ожиданием* дискретной случайной величины называют сумму произведений всех ее возможных значений на их вероятности. При этом должно удовлетворяться требование абсолютной сходимости ряда: $\sum_{i=1}^{\infty} x_i p_i$.

Формулу (3.13) можно распространить на случай непрерывных величин, заменив в ней отдельные значения x_i непрерывно изменяющейся величиной x, соответствующие вероятности p_i — элементом вероятности $f(x)\,dx$, сумму — интегралом.

Итак, для непрерывной случайной величины, возможные значения которой располагаются по всей оси Ox, математическое ожидание

$$M(x) = \int_{-\infty}^{\infty} x f(x)\,dx. \tag{3.14}$$

Математическое ожидание непрерывной случайной величины X, возможные значения которой принадлежат интервалу $[a, b]$,

$$M(x) = \int_a^b xf(x)\mathrm{d}x. \qquad (3.15)$$

Следует отметить, что встречаются случайные величины, для которых математического ожидания не существует.

Понятию математического ожидания случайной величины можно дать механическую интерпретацию. Распределение вероятностей дискретной случайной величины принимается за распределение масс, отнесенных к точкам $x_1, x_2, \ldots, x_n$ на оси абсцисс. Вся распределенная масса принимается равной единице. Тогда математическое ожидание, определяемое формулой (3.13), есть не что иное, как абсцисса центра тяжести всей системы материальных точек. Подобную же интерпретацию имеет математическое ожидание непрерывной случайной величины. В этом случае масса распределена по оси абсцисс непрерывно с плотностью $f(x)$.

Приведем без доказательств некоторые свойства математического ожидания.

1°. *Математическое ожидание постоянной величины равно самой постоянной:*

$$M(c) = c.$$

2°. *Постоянный множитель случайной величины может быть вынесен за знак математического ожидания:*

$$M(cx) = cM(x).$$

3°. *Математическое ожидание суммы двух случайных величин равно сумме их математических ожиданий:*

$$M[x + y] = M(x) + M(y).$$

Методом математической индукции это свойство можно распространить на произвольное число слагаемых величин. Математическое ожидание суммы нескольких случайных величин равно сумме их математических ожиданий:

$$M\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} M(x_i).$$

Следует отметить, что приведенное выше свойство справедливо как для независимых, так и для зависимых случайных величин.

4°. *Математическое ожидание произведения двух независимых случайных величин равно произведению их математических ожиданий:*

$$M[xy] = M(x)\,M(y).$$

Методом математической индукции это свойство можно распространить на произвольное число перемножаемых взаимно независимых случайных величин:

$$M\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} M(x_i).$$

Таким образом, математическое ожидание произведения нескольких взаимно независимых случайных величин равно произведению их математических ожиданий.

5°. *Математическое ожидание отклонения случайной величины от ее математического ожидания равно нулю:*

$$M[X - M(x)] = 0.$$

Пример 3.3. Найти математическое ожидание числа бракованных изделий в выборке из пяти изделий, если случайная величина X (число бракованных изделий) задана рядом распределения, построенном в примере 3.1.

Решение. По формуле (3.13) находим $M(x) = 0 \cdot 0{,}2373 + 1 \times \times 0{,}3955 + 2 \cdot 0{,}2637 + 3 \cdot 0{,}0879 + 4 \cdot 0{,}0146 + 5 \cdot 0{,}0010 = 1{,}25$.

Пример 3.4. Распределение содержания кремния в отливках из чугуна при определенном составе шихты таково:

Si, %	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
P	0,32	0,25	0,14	0,12	0,08	0,05	0,02	0,01	0,01

Определить математическое ожидание содержания кремния в отливках для данного состава шихты.

Решение. Случайной величиной X по условию задачи является содержание кремния в отливках из чугуна. По формуле (3.13) находим $M(x) = 1{,}2 \cdot 0{,}32 + 1{,}3 \cdot 0{,}25 + 1{,}4 \cdot 0{,}14 + 1{,}5 \cdot 0{,}12 + 1{,}6 \cdot 0{,}08 + + 1{,}7 \cdot 0{,}05 + 1{,}8 \cdot 0{,}02 + 1{,}9 \cdot 0{,}01 + 2{,}0 \cdot 0{,}01 = 1{,}373\%$.

К характеристикам положения случайной величины кроме математического ожидания относится также мода и медиана.

Модой M_0 дискретной случайной величины называется ее значение, имеющее наибольшую вероятность.

Модой M_0 *непрерывной случайной величины* называется такое ее значение, при котором плотность распределения имеет максимум. Геометрически моду можно интерпретировать как абсциссу точки максимума кривой распределения. Бывают *двухмодальные* и *многомодальные* распределения. Встречаются распределения, которые имеют минимум, но не имеют максимума. Такие распределения называются *антимодальными*.

Медианой случайной величины M_e называется такое ее значение, для которого справедливо равенство

$$P(X < M_e) = P(X > M_e),$$

т. е. равновероятно, что случайная величина окажется меньше или больше медианы. С геометрической точки зрения, медиана — это абсцисса точки, в которой площадь, ограниченная кривой распределения, делится пополам. Так как вся площадь, ограниченная кривой распределения и осью абсцисс, равна единице, то функция распределения в точке, соответствующей медиане, равна 0,5:

$$F(M_e) = P(X < M_e) = 0{,}5.$$

Следует отметить, что если распределение одномодальное и симметричное, то математическое ожидание, мода и медиана совпадают.

Дисперсия и среднее квадратическое отклонение. С помощью этих характеристик можно судить о рассеянии случайной величины относительно математического ожидания. Из свойства 5° математического ожидания очевидно, что выбор в качестве меры рассеяния математического ожидания отклонения случайной величины от ее математического ожидания не дает никакого результата. Поэтому в качестве меры рассеяния случайной величины берут математическое ожидание квадрата отклонения случайной величины от ее математического ожидания, которое называют *дисперсией случайной величины* X и обозначают $\sigma^2[X]$, а также σ_x^2 или $D[X]$:

$$\sigma_x^2 = M[X - M(x)]^2.$$

Для дискретной случайной величины дисперсия равна сумме произведений квадратов отклонений значений случайной величины от ее математического ожидания на соответствующие вероятности:

$$\sigma_x^2 = \sum_{i=1}^{n} [x_i - M(x)]^2 p_i. \qquad (3.16)$$

Для непрерывной случайной величины, распределение которой задано в виде плотности вероятности $f(x)$, дисперсия

$$\sigma_x^2 = \int_{-\infty}^{\infty} [x - M(x)]^2 f(x)\,dx. \qquad (3.17)$$

Недостатком дисперсии является то, что она имеет размерность квадрата случайной величины и ее неудобно использовать для характеристики разброса.

Этих недостатков лишено среднее квадратическое отклонение случайной величины, которое представляет собой положительный квадратный корень из дисперсии:

$$\sigma_x = \sqrt{\sigma_x^2}.$$

Рассмотрим некоторые свойства дисперсии.

1°. *Дисперсия постоянной величины равна нулю:*

$$\sigma^2(c) = 0.$$

Доказательство. Используя определение дисперсии и первое свойство математического ожидания, имеем

$$\sigma^2(c) = M[c - M(c)]^2 = M[c - c]^2 = M(0) = 0. \quad \square$$

2°. *Постоянный множитель случайной величины можно выносить за знак дисперсии, предварительно возведя его в квадрат:*

$$\sigma^2[cX] = c^2\sigma_x^2. \quad \square$$

Доказательство. Используя определение дисперсии и свойство 2° математического ожидания, имеем

$$\sigma^2[cX] = M[cX - M(x)]^2 = M[cX - cM(x)]^2 =$$
$$= M\{c^2[X - M(x)]^2\} = c^2 M[X - M(x)]^2 = c^2\sigma_x^2.$$

3°. *Дисперсия суммы двух независимых случайных величин равна сумме дисперсий этих величин:*

$$\sigma^2[X + Y] = \sigma_x^2 + \sigma_y^2.$$

Доказательство. Используя определение дисперсии и свойство математического ожидания $M(X+Y) =$

$=M(x)+M(y)$, дисперсию случайной величины $X+Y$ можно выразить следующим образом:

$$\sigma^2[X+Y]=M\{[X+Y-M[X+Y]\}^2=$$
$$=M[X-M(x)+Y-M(y)]^2.$$

Представим выражение в квадратных скобках в виде двучлена и запишем квадрат его суммы. Используя свойства математического ожидания суммы нескольких величин и произведения двух независимых случайных величин, получаем

$$\sigma^2[X+Y]=M[X-M(x)]^2+2M[X-M(x)]\times$$
$$\times M[Y-M(y)]+M[Y-M(y)]^2.$$

Согласно свойству 5° математического ожидания, второе слагаемое в последнем выражении равно нулю, следовательно,

$$\sigma^2[X+Y]=\sigma_x^2+\sigma_y^2. \quad \square$$

Методом математической индукции можно доказать это свойство для произвольного числа взаимно независимых случайных величин:

$$\sigma^2(\Sigma X_i)=\sum_{i=1}^{n}\sigma^2[X_i].$$

4°. *Дисперсия разности двух независимых случайных величин равна сумме их дисперсий:*

$$\sigma^2[X-Y]=\sigma_x^2+\sigma_y^2.$$

Доказательство. На основании свойства 3° дисперсии можно записать

$$\sigma^2[X+(-Y)]=\sigma_x^2+\sigma^2[-Y].$$

Согласно свойству 2° дисперсии имеем

$$\sigma^2[-Y]=(-1)^2\sigma_y^2=\sigma_y^2.$$

Окончательно получаем

$$\sigma^2[X-Y]=\sigma_x^2+\sigma_y^2. \quad \square$$

5°. *Дисперсия случайной величины равна разности между математическим ожиданием квадрата случайной величины X и квадратом ее математического ожидания.*

$$\sigma_x^2=M(x^2)-[M(x)]^2. \qquad (3.18)$$

Доказательство. Используя определение дисперсии и свойства математического ожидания, имеем

$$\sigma_x^2 = M[X - M(x)]^2 = M\{X^2 - 2XM(x) + [M(x)]^2\} =$$

$$= M(x^2) - 2M(x)M(x) + [M(x)]^2 = M(x^2) - [M(x)]^2. \quad \square$$

6°. *Дисперсия произведения двух независимых случайных величин X и Y определяется по формуле*

$$\sigma^2[XY] = \sigma_x^2\sigma_y^2 + [M(x)]^2\sigma_y^2 + [M(y)]^2\sigma_x^2.$$

Доказательство. Используя определение дисперсии и свойства математического ожидания, имеем

$$\sigma^2[XY] = M[XY - M(XY)]^2 = M[X^2Y^2] -$$

$$- 2M[XY]M[XY] + [M[XY]]^2 = M[X^2Y^2] - [M(XY)]^2.$$

При независимых X и Y величины X^2 и Y^2 также независимы, следовательно, $M[X^2Y^2] = M(x)^2M(y^2)$. Математические ожидания квадратов случайных величин X и Y можно записать в виде $M(x^2) = \sigma_x^2 + [M(x)]^2$; $M(y^2) = \sigma_y^2 + [M(y)]^2$. Окончательно имеем

$$\sigma^2[XY] = M(x^2)M(y^2) - [M(x)M(y)]^2 =$$

$$= \sigma_x^2\sigma_y^2 + [M(x)]^2\sigma_y^2 + [M(y)]^2\sigma_x^2.$$

Пример 3.5. Вычислить дисперсию числа бракованных изделий для распределения, приведенного в примере 3.1.

Решение. По определению дисперсии имеем $\sigma_x^2 = (0-1{,}25)^2 \times 0{,}2373 + (1-1{,}25)^2 \cdot 0{,}3955 + (2-1{,}25)^2 \cdot 0{,}2637 + (3-1{,}25)^2 \cdot 0{,}0879 + (4-1{,}25)^2 \cdot 0{,}0146 + (5-1{,}25)^2 \cdot 0{,}0010 = 0{,}938$; $M(x) = 1{,}25$ (см. пример 3.3).

Следует отметить, что дисперсию удобнее вычислять по формуле (3.18).

Пример 3.6. Вычислить дисперсию и среднее квадратическое отклонение содержания кремния в отливках из чугуна для распределения, приведенного в примере 3.4.

Решение. Найдем математическое ожидание квадрата случайной величины: $M(x^2) = 1{,}2^2 \cdot 0{,}32 + 1{,}3^2 \cdot 0{,}25 + 1{,}4^2 \cdot 0{,}14 + 1{,}5^2 \cdot 0{,}12 + 1{,}6^2 \cdot 0{,}08 + 1{,}7^2 \cdot 0{,}05 \times 1{,}8^2 \cdot 0{,}02 + 1{,}9^2 \cdot 0{,}01 + 2{,}0^2 \cdot 0{,}01 = 1{,}9179$ $[\%]^2$; $\sigma_x^2 = M(x^2) - [M(x)]^2 = 1{,}9179 - [1{,}373]^2 = 0{,}0328$ $[\%]^2$; $\sigma_x = 0{,}1811\%$.

Обобщением основных числовых характеристик случайной величины является понятие моментов случайной величины.

В гл. 1 было дано понятие моментов вариационных рядов. Сформулируем теперь определение начальных и центральных моментов как характеристик случайной величины.

Начальным моментом q-го порядка случайной величины называют математическое ожидание величины X^q:

$$\nu_q = M[X^q].$$

Начальный момент дискретной случайной величины

$$\nu_q = \sum_{i=1}^{n} x_i^q p_i;$$

начальный момент непрерывной случайной величины

$$\nu_q = \int_{-\infty}^{+\infty} x^q f(x)\,dx.$$

Центральным моментом q-го порядка случайной величины называют математическое ожидание величины $[X-M(x)]^q$:

$$\mu_q = M\{[X - M(x)]^q\}.$$

Центральный момент дискретной случайной величины

$$\mu_q = \sum_{i=1}^{n} [x_i - M(x)]^q p_i;$$

центральный момент непрерывной случайной величины

$$\mu_q = \int_{-\infty}^{\infty} [x - M(x)]^q f(x)\,dx.$$

Для начальных и центральных моментов случайной величины справедливы те же соотношения, что и для моментов вариационного ряда.

Начальный момент первого порядка представляет собой математическое ожидание, а центральный момент второго порядка — дисперсию случайной величины.

Нормированный центральный момент третьего порядка служит характеристикой скошенности или асимметрии распределения (коэффициент асимметрии):

$$A = \mu_3/\sigma^3.$$

Нормированный центральный момент четвертого порядка служит характеристикой островершинности или плосковершинности распределения (эксцесс):

$$E = \mu_4/\sigma^4 - 3.$$

Пример 3.7. Случайная величина X задана плотностью распределения:

$$f(x)=\begin{cases}0 & \text{при } x<0,\\ ax^2 & \text{при } 0<x<2,\\ 0 & \text{при } x>2.\end{cases}$$

Найти коэффициент a, интегральную функцию, математическое ожидание, дисперсию, асимметрию и эксцесс.

Р е ш е н и е. Площадь, ограниченная кривой распределения, численно равна

$$\int_0^2 f(x)\,dx=\int_0^2 ax^2\,dx=\frac{ax^3}{3}\Big|_0^2=\frac{8}{3}a.$$

Учитывая, что эта площадь должна быть равна единице, находим $a=3/8$. Для определения интегральной функции используем формулу (3.11):

$$F(x)=\int_0^x f(x)\,dx=\int_0^x 3/8\,x^2\,dx=\int_0^x 3/8\,x^3/3=x^3/8,$$

следовательно,

$$F(x)=\begin{cases}0 & \text{при } x<0,\\ x^3/8 & \text{при } 0<x<2,\\ 1 & \text{при } x<2.\end{cases}$$

По формуле (3.15) найдем математическое ожидание:

$$M(x)=\int_0^2 xf(x)\,dx=\int_0^2 3/8\,x^3\,dx=3/8\Big|_0^2 x^4/4=1,5.$$

Дисперсию определим по формуле (3.18). Для этого найдем сначала математическое ожидание квадрата случайной величины:

$$M(x^2)=\int_0^2 x^2f(x)\,dx=\int_0^2 3/8\,x^4\,dx=3/8\Big|_0^2 x^5/5=2,4$$

Таким образом,

$$\sigma_x^2=M(x^2)-[M(x)]^2=2,4-(1,5)^2=0,15.$$

Центральные моменты третьего и четвертого порядков вычислим через начальные моменты. Имеем:

$$\nu_1=M(x)=1,5;\quad \nu_2=M(x^2)=2,4;$$

$$\nu_3=M(x^3)=\int_0^2 x^3f(x)\,dx=\int_0^2 3/8\,x^5\,dx=3/8\Big|_0^2 x^6/6=4;$$

$$\nu_4=M(x^4)=\int_0^2 x^4f(x)\,dx=\int_0^2 3/8\,x^6\,dx=3/8\Big|_0^2 x^7/7=6,8571;$$

$$\mu_3 = \nu_3 - 3\nu_1\,\nu_2 + 2\nu_1^3 = 4 - 3\cdot 1{,}5\cdot 2{,}4 + 2\cdot 1{,}5^3 = 0{,}05; -$$

$$\mu_4 = \nu_4 - 4\nu_1\,\nu_3 + 6\nu_1^2\,\nu_2 - 3\nu_1^4 = 6{,}8571 - 4\cdot 1{,}5\cdot 4 +$$

$$+ 6\cdot 1{,}5^2\cdot 2{,}4 - 3\cdot 1{,}5^4 = 0{,}0696;$$

$$A = \mu_3/\sigma^3 = -0{,}05/(0{,}3873)^3 = -0{,}86;$$

$$E = \mu_4/\sigma^4 - 3 = 0{,}0696/(0{,}3873)^4 - 3 = -0{,}093.$$

Введем понятие центрированной и нормированной случайной величины. Отклонение случайной величины X от ее математического ожидания $M(x)$, т. е. $\overset{\circ}{X} = X - M(x)$, называют *центрированной* случайной величиной.

Математическое ожидание центрированной случайной величины $M[\overset{\circ}{X}]$ равно нулю, а дисперсия $\sigma^2[\overset{\circ}{X}]$ равна дисперсии случайной величины X.

Доказательство. На основании свойства 5° математического ожидания

$$M[\overset{\circ}{X}] = M[X - M(x)] = 0.$$

Используя свойства дисперсии 1° и 4°, получаем

$$\sigma^2[\overset{\circ}{X}] = \sigma^2[X - M(x)] = \sigma_x^2. \quad \square$$

Центрирование случайной величины геометрически равносильно переносу начала координат в точку, абсцисса которой равна математическому ожиданию.

Нормированной случайной величиной (T) называют центрированную случайную величину, выраженную в долях среднего квадратического отклонения:

$$T = \overset{\circ}{X}/\sigma_x = (X - M(x))/\sigma_x.$$

Математическое ожидание нормированной случайной величины $M[T]$ равно нулю, а дисперсия $\sigma^2[T]$ равна единице.

Доказательство. На основании свойств 2° и 5° математического ожидания имеем

$$M[T] = M\left[\frac{X - M(x)}{\sigma_x}\right] = \frac{1}{\sigma_x} M[X - M(x)] = 0.$$

Используя свойства дисперсии, получим

$$\sigma^2[T] = \sigma^2\left[\frac{X - M(x)}{\sigma_x}\right] = \frac{1}{\sigma_x^2}\sigma^2[X - M(x)] = \sigma_x^2/\sigma_x^2 = 1. \quad \square$$

§ 3.3. Равномерное распределение

Непрерывная случайная величина X имеет равномерное распределение на интервале $[a, b]$, *если на этом интервале плотность распределения случайной величины постоянна, а вне его равна нулю, т. е. если*

$$f(x)=\begin{cases}0 & \text{при } x<a,\\ c & \text{при } a\leqslant x\leqslant b,\\ 0 & \text{при } x>b,\end{cases}$$

где c=const. Равномерное распределение иногда называются *законом равномерной плотности.*

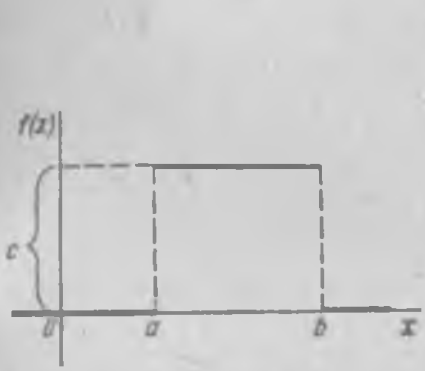

Рис. 3.13

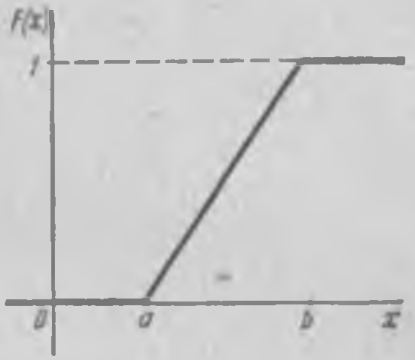

Рис. 3.14

График плотности $f(x)$ равномерного распределения изображен на рис. 3.13. Найдем значение постоянной c. Так как площадь, ограниченная кривой распределения, равна единице и все значения случайной величины принадлежат интервалу $[a, b]$, то должно выполняться равенство

$$\int_a^b f(x)\,dx=1, \text{ или } \int_a^b c\,dx=1,$$

отсюда $c=1/(b-a)$.

Итак, закон равномерного распределения аналитически можно записать в виде

$$f(x)=\begin{cases}0 & \text{при } x<a,\\ 1/(b-a) & \text{при } a\leqslant x\leqslant b,\\ 0 & \text{при } x>b.\end{cases}$$

Найдем выражение функции распределения $F(x)$ для равномерного распределения на интервале $[a, b]$. Согласно формуле (3.11),

$$F(x)=\int_a^x f(x)\,dx=\int_a^x \frac{1}{b-a}\,dx=\frac{x}{b-a}\Bigg|_a^x=\frac{x-a}{b-a}.$$

При $x<a$ имеем $F(x)=0$, $F(x)=1$ при $x>b$.

Таким образом,

$$F(x)=\begin{cases}0 & \text{при } x<a,\\ (x-a)/(b-a) & \text{при } a\leqslant x\leqslant b,\\ 1 & \text{при } x>b.\end{cases}$$

График функции $F(x)$ изображен на рис. 3.14.

Определим основные числовые характеристики случайной величины X, имеющей равномерное распределение. Математическое ожидание

$$M(x)=\int_a^b \frac{x}{b-a}\,dx=\frac{x^2}{2(b-a)}\Bigg|_a^b=\frac{a+b}{2}.$$

Итак, математическое ожидание равномерного распределения находится посередине интервала $[a, b]$. В силу симметричности распределения медиана величины X совпадает с математическим ожиданием: $M_e=(a+b)/2$. Моды равномерное распределение не имеет. Дисперсию случайной величины X находим по формуле (3.17):

$$\sigma^2=\mu_2=\int_a^b\left(x-\frac{(b+a)}{2}\right)^2\frac{1}{b-a}\,dx=\frac{(b-a)^2}{12}.$$

Среднее квадратическое отклонение $\sigma=(b-a)/(2\sqrt{3})$.

В силу симметричности распределения центральный момент третьего порядка, а следовательно, и коэффициент асимметрии для закона равномерной плотности равны нулю: $A=\mu_3/\sigma^2=0$. Четвертый центральный момент

$$\mu_4=\int_a^b\left(x-\frac{b+a}{2}\right)^4\frac{1}{b-a}\,dx=\frac{(b-a)^4}{80}.$$

Коэффициент эксцесса $E=\mu_4/\sigma^4-3=-1{,}2$.

Вероятность попадания случайной величины X, имеющей равномерное распределение, на участок $[\alpha, \beta]$,

который является частью участка $[a, b]$, определяется по формуле

$$P(\alpha<X<\beta)=\int_{\alpha}^{\beta}\frac{dx}{b-a}=\frac{\beta-\alpha}{b-a}.$$

§ 3.4. Нормальное распределение

Нормальное распределение — наиболее часто встречающийся вид распределения. С ним приходится сталкиваться при анализе погрешностей измерений, контроле технологических процессов и режимов, а также при анализе и прогнозировании различных явлений в биологии, медицине и других областях знаний.

Анализ общих условий возникновения нормального распределения показывает, что наиболее важным условием является формирование признака как суммы большого числа взаимно независимых слагаемых, ни одно из которых не характеризуется исключительно большой по сравнению с другими дисперсией. В производственных условиях такие предпосылки в основном соблюдаются. Рассмотрим, например, обработку деталей на станке-автомате. Отклонение размеров деталей от номинального вызывается многими причинами, более или менее независимыми друг от друга. К ним относятся: колебание режима обработки, неточность установки и базировки деталей в приспособлении, неоднородность обрабатываемого материала, неточность изготовления и износ режущего инструмента, упругие деформации узлов станка под влиянием усилий обработки, износ деталей станка и т. д. Каждая из этих причин влияет на размер детали, так что то отклонение, которое фактически регистрируется при измерениях, является суммой большого числа отклонений. Если слагаемые отклонения примерно одного порядка, то суммарное отклонение является случайной величиной с нормальным распределением.

Термин «нормальное распределение» применяется в условном смысле как общепринятый в литературе термин, хотя и не совсем удачный. Так, утверждение, что какой-то признак подчиняется нормальному закону, вовсе не означает наличие каких-то незыблемых норм, якобы лежащих в основе явления, отражением которого является рассматриваемый признак, а подчинение дру-

гим видам законов не означает какую-то анормальность данного явления.

Главная особенность нормального распределения состоит в том, что оно является предельным, к которому приближаются другие распределения. Нормальное распределение впервые открыто Муавром в 1733 г.* Нормальному закону распределения подчиняются только непрерывные случайные величины. Поэтому распределение

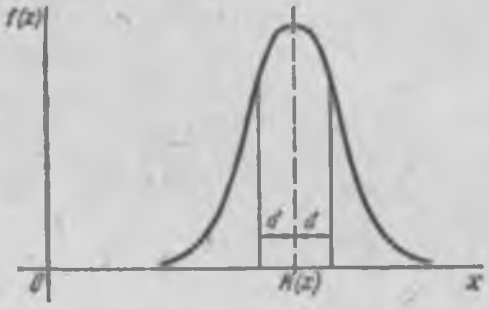

Рис. 3.15

нормальной совокупности может быть задано в виде плотности распределения:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-a)^2/(2\sigma^2)} \qquad (3.19)$$

или в виде функции распределения:

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x-a)^2/(2\sigma^2)}\,dx. \qquad (3.20)$$

График плотности нормального распределения называют *нормальной кривой* (рис. 3.15). Она представляет собой колоколообразную фигуру, симметричную относительно прямой, проходящей через точку $x=a$, и асимптотически приближающуюся к оси абсцисс при $x \to \pm\infty$. Как следует из аналитического выражения плотности, нормальное распределение определяется параметрами a и σ. На рис. 3.4 изображен график интегральной функции нормального распределения.

* Нормальное распределение часто называют *законом Гаусса — Лапласа* по имени математиков, открывших этот закон независимо от работ Муавра.

Найдем математическое ожидание и дисперсию случайной величины, подчиняющейся нормальному закону, для чего используем общие определения числовых характеристик случайных величин:

$$M(x)=\int\limits_{-\infty}^{\infty} xf(x)\,dx=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{\infty} xe^{-(x-a)^2/(2\sigma^2)}\,dx.$$

Введем новую переменную $t=(x-a)/\sigma$. Имеем: $x=\sigma t+a$; $dx=\sigma dt$. Нетрудно убедиться, что новые пределы интегрирования совпадают со старыми. Таким образом,

$$M(x)=\frac{a}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}(\sigma t+a)e^{-t^2/2}\,dt=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}\sigma t\,e^{-t^2/2}\,dt+$$

$$+\frac{a}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}e^{-t^2/2}\,dt=a.$$

Первое слагаемое в правой части равенства равно нулю, так как под знаком интеграла с симметричными пределами находится нечетная функция. Второе слагаемое равно a, поскольку

$$\int\limits_{-\infty}^{\infty}e^{-t^2/2}dt=\sqrt{2\pi}\,.$$

Итак, $M(x)=a$.

Найдем дисперсию случайной величины, распределенной нормально:

$$\sigma_x^2=\int\limits_{-\infty}^{\infty}(x-a)^2f(x)\,dx=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}(x-a)^2e^{-(x-a)^2/(2\sigma^2)}\,dx.$$

Вводя в это выражение новую переменную $t=(x-a)/\sigma$ и принимая во внимание, что новые пределы интегрирования равны старым, получаем

$$\sigma_x^2=\frac{\sigma^2}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}t^2e^{-t^2/2}\,dt=\frac{\sigma^2}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}t\cdot te^{-t^2/2}\,dt.$$

Интегрируем по частям, положив $u=t$, $te^{-t^2/2}\,dt=dv$, откуда $du=dt$, $v=\int\limits_{-\infty}^{\infty}te^{-t^2/2}\,dt$. Окончательно имеем $\sigma_x^2=\sigma^2$.

Таким образом, становится ясен статистический смысл параметров a и σ нормальной кривой распределения.

Рассмотрим некоторые свойства нормального распределения.

1°. *Функция плотности нормального распределения определена на всей оси Ox, т. е. каждому значению x соответствует вполне определенное значение функции.*

2°. *При всех значениях x (как положительных, так и отрицательных) функция плотности принимает положительные значения, т. е. нормальная кривая расположена над осью Ox.*

3°. *Предел функции плотности при неограниченном возрастании x равен нулю:*

$$\lim_{x \to \pm\infty} f(x) = 0.$$

Графически это состоит в том, что ветви кривой асимптотически приближаются к оси *Ox*.

4°. *Функция плотности нормального распределения в точке $x=a$ имеет максимум, равный*

$$f(a) = \frac{1}{\sigma\sqrt{2\pi}}.$$

В этом нетрудно убедиться, найдя первую производную от $f(x)$:

$$f'(x) = -\frac{x-a}{\sigma^3\sqrt{2\pi}}\, e^{-(x-a)^2/(2\sigma^2)}.$$

Так как первая производная при переходе через точку, в которой она обращается в нуль, меняет знак от «+» к «—», то имеем в точке $x=a$ максимум $f(x)$.

5°. *График функции плотности $f(x)$ симметричен относительно прямой*, проходящей через точку $x=a$.

Отсюда следует равенство для нормально распределенной величины моды, медианы и математического ожидания.

6°. *Кривая распределения имеет две точки перегиба с координатами* $(a-\sigma;\ 1/(\sigma\sqrt{2\pi e}))$ *и* $(a+\sigma;\ 1/(\sigma\sqrt{2\pi e}))$.

Для отыскания координат этих точек найдем вторую производную:

$$f''(x) = \frac{1}{\sigma^3\sqrt{2\pi}}\, e^{-(x-a)^2/(2\sigma^2)}\left[\frac{(x-a)^2}{\sigma^2} - 1\right].$$

При $x=a+\sigma$ и $x=a-\sigma$ вторая производная обращается в нуль, а при переходе через эти точки она меняет знак. Подставляя эти значения аргумента в выражение функции плотности, находим ординаты точек перегиба.

7°. *Нечетные центральные моменты нормального распределения равны нулю.*

Используя определение центральных моментов, можно вывести следующее рекуррентное соотношение:

$$\mu_q = (q-1)\sigma^2 \mu_{q-2}.$$

По этой формуле можно убедиться, что центральные моменты нечетных порядков равны нулю, поскольку $\mu_1=0$, а для центральных моментов четного порядка получить следующие выражения: $\mu_2=\sigma^2$, $\mu_4=3\sigma^4$, $\mu_6=15\sigma^6$.

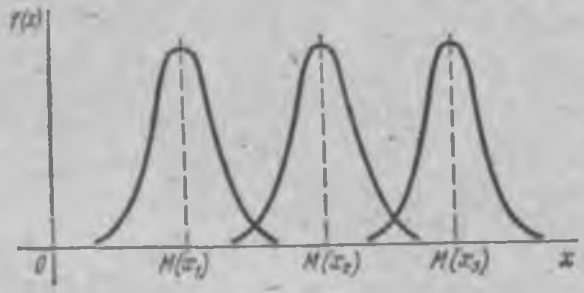

Рис. 3.16

8°. *Коэффициенты асимметрии и эксцесса нормального распределения равны нулю.* Действительно,

$$A = \mu_3/\sigma^3 = 0;\ E = \mu_4/\sigma^4 - 3 = 0.$$

Становится ясной важность вычисления этих коэффициентов для эмпирических рядов распределения, так как они характеризуют скошенность и крутость данного ряда по сравнению с нормальным.

9°. *Форма нормальной кривой не изменяется при изменении величины параметра a* (математического ожидания). При уменьшении или увеличении математического ожидания график кривой сдвигается влево или вправо (рис. 3. 16).

При изменении же параметра σ меняется форма кривой. С возрастанием σ максимальная ордината кривой распределения уменьшается, а с уменьшением σ — возрастает. Согласно свойству плотности распределения, площадь, ограниченная кривой распределения и осью

абсцисс, равна единице. Поэтому с ростом σ кривая распределения сжимается к оси абсцисс и растягивается вдоль нее; с уменьшением σ нормальная кривая растягивается вдоль оси ординат (рис. 3.17).

Плотность распределения с параметрами $a=0$, $\sigma=1$ называют *нормированной плотностью,* а ее график — *нормированной нормальной кривой.* Нормированную нормальную кривую можно представить как кривую распределения нормированной случайной величины

$$T = (X - M(x))/\sigma.$$

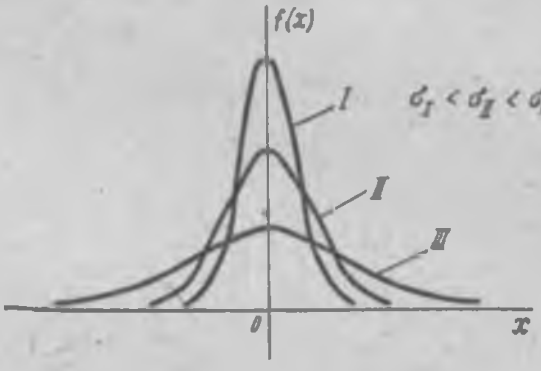

Рис. 3.17

Нормированную плотность распределения используют для расчета теоретической кривой распределения, соответствующей данному эмпирическому ряду. Для этого переменные заменяют выражением $t=(x-a)/\sigma$. Значения нормированной плотности $f(t)=(1/\sqrt{2\pi})e^{-t^2/2}$ для различных t приведены в табл. 1 приложений. Следует иметь в виду, что нормированная плотность нормального распределения является четной функцией, т. е. $f(-t)=f(t)$.

На практике часто необходимо вычислять вероятность того, что нормально распределенная случайная величина находится в определенных границах. Расчет легко провести с помощью значений специальной функции (*функции Лапласа* или *интеграла вероятности*):

$$\Phi(t) = \frac{2}{\sqrt{2\pi}} \int_0^t e^{-t^2/2}\, dt.$$

Эти значения приведены в табл. 2 приложений. Функция Лапласа — нечетная функция, т. е. $\Phi(-t)=-\Phi(t)$.

В § 3.1 показано, что вероятность того, что любая случайная величина X примет значение, принадлежащее интервалу $[\alpha, \beta]$, можно выразить через функцию распределения $P(\alpha<X<\beta)=F(\beta)-F(\alpha)$.

Перейдем к нормированной случайной величине $T=[X-M(x)]/\sigma$. Неравенства $\alpha<X<\beta$ и $[\alpha-M(x)]/\sigma<[X-M(x)]/\sigma<[\beta-M(x)]/\sigma$ равносильны. Поэтому вероятности этих неравенств равны между собой:

$$P(\alpha<X<\beta)=P\left\{\frac{\alpha-M(x)}{\sigma}<\frac{X-M(x)}{\sigma}<\frac{\beta-M(x)}{\sigma}\right\}.$$

Используя предыдущее равенство, запишем

$$P(t_1<T<t_2)=F(t_2)-F(t_1), \qquad (3.21)$$

где $t_1=(\alpha-M(x))/\sigma$; $t_2=(\beta-M(x))/\sigma$. Считая, что нормированная случайная величина T имеет нормальное распределение и используя определение интегральной функции распределения, получаем

$$F(t)=P(T<t)=\int_{-\infty}^{t} f(t)\,dt=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-t^2/2}dt=$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} e^{-t^2/2}dt+\frac{1}{\sqrt{2\pi}}\int_{0}^{t} e^{-t^2/2}\,dt.$$

Первое слагаемое численно равно половине площади под нормированной нормальной кривой распределения; второе слагаемое является выражением функции Лапласа. Итак, для нормально распределенной случайной величины

$$F(t)=1/2+1/2\,\Phi(t). \qquad (3.22)$$

Учитывая последнее выражение, формулу (3.21) приведем к виду

$$P\{t_1<T<t_2\}=1/2\,\Phi(t_2)-1/2\,\Phi(t_1). \qquad (3.23)$$

Определим вероятность того, что нормально распределенная случайная величина X отклоняется от своего математического ожидания $M(x)$ на величину, меньшую ε, т. е. найдем вероятность осуществления неравенства $|X-M(x)|<\varepsilon$, или (что одно и то же) вероятность двойного неравенства $M(x)-\varepsilon<X<M(x)+\varepsilon$. Перейдем к нормированной случайной величине T, заменив границы

отклонения случайной величины X на нормированные: $t=(M(x)\pm\varepsilon-M(x))/\sigma=\pm\varepsilon/\sigma$. По формуле (3.23) получаем

$$P\{t_1<T<t_2\}=\tfrac{1}{2}\Phi(t_2)-\tfrac{1}{2}\Phi(t_1)=$$
$$=\tfrac{1}{2}\Phi(\varepsilon/\sigma)-\tfrac{1}{2}\Phi(-\varepsilon/\sigma)=\Phi(\varepsilon/\sigma).$$

Итак, окончательно имеем

$$P(|X-M(x)|<\varepsilon)=\Phi(\varepsilon/\sigma). \qquad (3.24)$$

В частности, при $M(x)=0$

$$P\{|X|<\varepsilon\}=\Phi(\varepsilon/\sigma).$$

Выразив отклонение случайной величины X в долях среднего квадратического отклонения, т. е. положив $\varepsilon=t\sigma$, равенство (3.24) можно записать в виде

$$P\{|X-M(x)|<t\sigma\}=\Phi(t). \qquad (3.25)$$

Таким образом, значение функции $\Phi(t)$ при заданном t определяет вероятность того, что отклонение нормально распределенной случайной величины по абсолютной величине меньше $t\sigma$.

При $t=1$ $P(|X-M(x)|<\sigma)=\Phi(1)=0{,}6827$; при $t=2$ $P(|X-M(x)|<2\sigma)=\Phi(2)=0{,}9545$; при $t=3$ $P(|X-M(x)|<3\sigma=\Phi(3)=0{,}9973$.

Из последнего равенства следует, что практически рассеяние нормально распределенной случайной величины заключено на участке $M(x)\pm3\sigma$. Вероятность того, что случайная величина X не попадет за этот участок, очень мала, а именно равна 0,0027, т. е. это событие может произойти лишь в трех случаях из 1000. Такие события можно считать практически невозможными. На приведенном рассуждении основано **правило трех сигм**, которое формулируется следующим образом: *если случайная величина имеет нормальное распределение, то отклонение этой величины от математического ожидания по абсолютной величине практически не превосходит утроенного среднего квадратического отклонения.*

Используя это правило, ориентировочно оценивают среднее квадратическое отклонение. Для этого из ряда наблюдения выбирают максимальное и минимальное значения и их разность делят на шесть. Полученное число является грубой оценкой среднего квадратического отклонения при условии, что распределение признака нормально.

Пример 3.8. Размер детали задан полем допуска 10—12 мм. Оказалось, что средний размер деталей равен 11,4 мм, а среднее квадратическое отклонение — 0,7 мм. Считая, что размер детали подчиняется закону нормального распределения, определить вероятность появления брака по заниженному и завышенному размеру.

Решение. По условию задачи, $M(x)=11{,}4$ мм, $\sigma=0{,}7$ мм. Нижняя граница поля допуска $T_н=10$ мм. Верхняя граница поля допуска $T_в=12$ мм. Браком по заниженному размеру является деталь с размером, выходящим за нижнюю границу допуска; браком по завышенному размеру — деталь с размером, выходящим за верхнюю границу поля допуска. Искомые вероятности соответственно равны:

$$P(X<T_н)=F(T_н)=1/2+1/2\,\Phi\left[\frac{T_н-M(x)}{\sigma}\right]=$$

$$=1/2+(1/2)\,\Phi(-2)=1/2-1/2\,\Phi(2)=0{,}0225;$$

$$P(X>T_в)=1-P(X<T_в)=1-F(T_в)=1/2-$$

$$-1/2\,\Phi\left[\frac{T_в-M(x)}{\sigma}\right]=1/2-1/2\,\Phi(0{,}86)=0{,}1950.$$

При решении использованы данные, приведенные в табл. 2 приложений.

Пример 3.9. Установлено, что при измерении диаметра валика микрометром случайная погрешность подчиняется нормальному закону со средним квадратическим отклонением, равным 0,12 мм. Систематической погрешностью измерения можно пренебречь. Найти вероятность того, что измерение производится с ошибкой, не превосходящей по абсолютной величине 0,2 мм.

Решение. По условию $\varepsilon=0{,}2$; $\sigma=0{,}12$ мм. Искомая вероятность

$$P(|X|<0{,}2)=\Phi(0{,}2/0{,}12)=\Phi(1{,}67)=0{,}9051.$$

Пример 3.10. В табл. 3.1—3.3 приведены три способа расчета теоретического нормального распределения по двум известным характеристикам x и s эмпирического ряда распределения диаметра валика после шлифовки, построенного в гл. 1 (табл. 1.4). При расчетах статистические характеристики эмпирического ряда распределения приравнивают числовым характеристикам теоретического:

$$M(x)=\bar{x}=6{,}7578;\ \sigma=s=0{,}0314.$$

В табл. 3.1. расчет произведен по нормированной плотности нормального распределения $f(t)=(1/\sqrt{2\pi})e^{-t^2/2}$, где $t=[x-M(x)]\sigma$ — нормированное отклонение случайной величины. В этом случае интервальный ряд заменяется дискретным в предположении, что частота (или частость) относится к серединам интервалов. После вычисления нормированных отклонений t находятся значения нормированной плотности $f(t)$ по табл. 1 приложений. Плотность распределения случайной величины X можно представить в виде

$$f(x)=\frac{1}{\sigma}\,\frac{1}{\sqrt{2\pi}}e^{-t^2/2}=\frac{1}{\sigma}f(t).$$

Таблица 3.1

Расчет теоретической кривой распределения диаметра валиков после шлифовки с помощью нормированной плотности нормального распределения $f(t)=(1/\sqrt{2\pi})\ e^{-t^2/2}$

Интервалы $a-b$	Эмпирическая частота $m_э$	$x=\frac{a+b}{2}$	$t=\frac{X-M(x)}{\sigma}$	$f(t)$	$\frac{h}{\sigma}f(t)$	$n\frac{h}{\sigma}f(t)$	Теоретическая частота $m_т$
6,67—6,69	2	6,68	—2,48	0,0184	0,0117	2,34	2
6,69—6,71	15	6,70	—1,84	0,0734	0,0467	9,35	9
6,71—6,73	17	6,72	—1,20	0,1942	0,1237	24,74	25
6,73—6,75	44	6,74	—0,57	0,3391	0,2160	43,19	43
6,75—6,77	52	6,76	0,07	0,3980	0,2535	50,70	51
6,77—6,79	44	6,78	0,71	0,3101	0,1975	39,50	40
6,79—6,81	14	6,80	1,34	0,1626	0,1036	20,71	21
6,81—6,83	11	6,82	1,98	0,0562	0,0358	7,16	7
6,83—6,85	1	6,84	2,62	0,0129	0,0082	1,64	2
Σ	200				0,9967		200

Таблица 3.2

Расчет теоретической кривой распределения диаметра валиков после шлифовки с помощью функции $P(a<X<b)=1/2\{\Phi(t_2)-\Phi(t_1)\}$

Интервалы $a-b$	Эмпирическая частота $m_э$	$t_1=\frac{a-M(x)}{\sigma}$	$t_2=\frac{b-M(x)}{\sigma}$	$\frac{1}{2}\Phi(t_1)$	$\frac{1}{2}\Phi(t_2)$	$P(a<X<b)$	nP	Теоретическая частота $m_т$
6,67—6,69	2	—2,80	—2,16	—0,4974	—0,4846	0,0128	2,56	3
6,69—6,71	15	—2,16	—1,52	—0,4846	—0,4357	0,0489	9,78	10
6,71—6,73	17	—1,52	—0,88	—0,4357	—0,3106	0,1251	25,02	25
6,73—6,75	44	—0,88	—0,25	—0,3106	—0,0987	0,2119	42,38	42
6,75—6,77	52	—0,25	0,39	—0,0987	0,1517	0,2504	50,08	50
6,77—6,79	44	0,39	1,03	0,1517	0,3485	0,1968	39,36	39
6,79—6,81	14	1,03	1,66	0,3485	0,4515	0,1030	20,60	21
6,81—6,83	11	1,66	2,30	0,4515	0,4892	0,0377	7,54	8
6,83—6,85	1	2,30	2,94	0,4892	0,4983	0,0091	1,82	2
Σ	200					0,9957		200

Таблица 3.3

Расчет интегральной кривой распределения диаметра валиков после шлифовки с помощью функции $F(b)=1/2+1/2\Phi(t)$

Интервалы $a-b$	Эмпирическая частость $\omega_э$	Накопленная эмпирическая частость $\omega_{нак}$	$t=\frac{b-M(x)}{\sigma}$	$\frac{1}{2}\Phi(t)$	$F(b)$
6,67—6,69	0,010	0,010	—2,16	—0,4846	0,0154
6,69—6,71	0,075	0,085	—1,52	—0,4357	0,0643
6,71—6,73	0,085	0,170	—0,88	—0,3106	0,1894
6,73—6,75	0,220	0,390	—0,25	—0,0987	0,4013
6,75—6,77	0,260	0,650	0,39	0,1517	0,6517
6,77—6,79	0,220	0,870	1,03	0,3485	0,8485
6,79—6,81	0,070	0,940	1,66	0,4515	0,9515
6,81—6,83	0,055	0,995	2,30	0,4892	0,9892
6,83—6,85	0,005	1,000	2,94	0,4983	0,9983
Σ	1,000				

Произведение $(1/\sigma)f(t)$ является теоретической частостью, приходящейся на середину интервала. Для определения частоты, распределенной по всей ширине интервала, эту величину следует умножить на ширину интервала h и на общее число наблюдений $n=\Sigma m_т$. При округлении вычисленных значений теоретических частот необходимо, чтобы соблюдалось равенство сумм эмпирических и теоретических частот, т. е. $\Sigma m_э=\Sigma m_т$. Из-за того, что расчет вероятностей ведется по серединам интервалов, этот способ является недостаточно точным.

При втором способе расчета теоретической кривой нормального распределения вероятность попадания случайной величины в каждый интервал распределения вычисляется по формуле

$$P(a<X<b)=1/2\{\Phi(t_2)-\Phi(t_1)\},$$

где $t_1=[a-M(x)]/\sigma$ и $t_2=[b-M(x)]/\sigma$ — нормированные отклонения начала и конца интервалов.

По табл. 2 приложений находим для вычисленных нормированных отклонений значения функции распределения $\Phi(t)$. Соответствующие вычисления приведены в табл. 3.2. В результате получаем теоретическую частоту признака, относящуюся ко всей ширине интервала.

Накопленные теоретические частоты (или частости) можно вычислить с помощью интегральной функции $F(b)=1/2+1/2\Phi(t)$, где t — нормированное отклонение верхних границ интервалов: $t=[b-M(x)]/\sigma$. В итоге получаем накопленную теоретическую частость, отнесенную к верхним границам интервалов (табл. 3.3).

Теоретическая нормальная кривая и интегральная кривая построены на эмпирических графиках ряда распределения (гистограмме и кумуляте) на рис. 3.18 и 3.19.

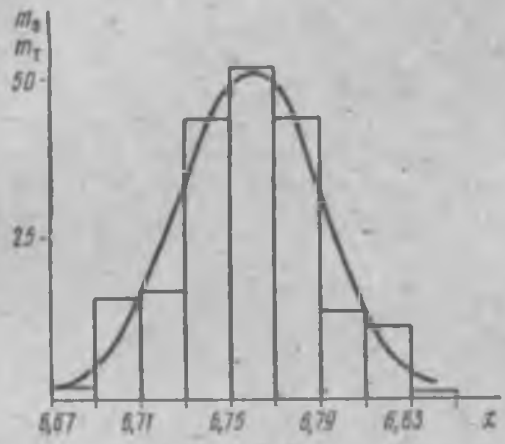

Рис. 3.18

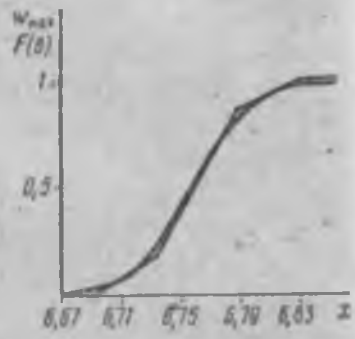

Рис. 3.19

§ 3.5. Биномиальное распределение

Биномиальным является распределение вероятностей появления m числа событий в n независимых испытаниях, в каждом из которых вероятность появления события постоянна и равна p. Вероятность возможного числа появления события вычисляется по формуле Бернулли

$$P(X = m) = C_n^m p^m q^{n-m} \quad (m = 0, 1, \ldots, n),$$

где $q=1-p$. Постоянные n и p, входящие в это выражение, — параметры биномиального закона.

Биномиальным распределением описывается распределение вероятностей только дискретной случайной величины. Возможными значениями случайной величины X являются $m=0,1, \ldots, n$. Биномиальному распределению подчиняется, например, число бракованных изделий в выборках из неограниченной партии продукции. Биномиальное распределение может быть задано в виде ряда:

$X=m$	0	1	2	...	l	...	n
$P_{m,n}$	$C_n^0 p^0 q^n$	$C_n^1 p^1 q^{n-1}$	$C_n^2 p^2 q^{n-2}$	...	$C_n^l p^l q^{n-l}$	...	$C_n^n p^n q^0$

и в виде функции распределения

$$f(x)=\begin{cases}0 & \text{при } m\leqslant 0,\\ \sum\limits_{m_i<m} P_{m,n} & \text{при } 0<m\leqslant n,\\ 1 & \text{при } m>n.\end{cases}$$

Найдем числовые характеристики случайной величины, подчиняющейся биномиальному распределению. Запишем выражения начальных моментов первого и второго порядка:

$$\nu_1=\sum_{m=0}^{n} mC_n^m p^m q^{n-m}, \tag{3.26}$$

$$\nu_2=\sum_{m=0}^{n} m^2 C_n^m p^m q^{n-m}. \tag{3.27}$$

Для вычисления сумм в формулах (3.26)—(3.27) продифференцируем два раза по p выражение

$$(p+q)^n=\sum_{m=0}^{n} C_n^m p^m q^{n-m}.$$

В результате получаем:

$$n(p+q)^{n-1}=\sum_{m=0}^{n} mC_n^m p^{m-1} q^{n-m}; \tag{3.28}$$

$$n(n-1)(p+q)^{n-2}=\sum_{m=0}^{n} m(m-1)C_n^m p^{m-2} q^{n-m}. \tag{3.29}$$

Умножая левую и правую части равенства (3.28) на p, имеем

$$np(p+q)^{n-1}=\sum_{m=0}^{n} mC_n^m p^m q^{n-m}. \tag{3.30}$$

Правые части равенства (3.26) и (3.30) равны между собой, следовательно, равны и левые части, т. е.

$$\nu_1=np(p+q)^{n-1}.$$

Учитывая, что $p+q=1$, имеем

$$\nu_1=np \text{ или } M(x)=np. \tag{3.31}$$

Итак, *математическое ожидание числа появлений события в n независимых испытаниях равно произведению числа испытаний на вероятность появления события в каждом испытании.*

Умножим обе части равенства (3.29) на p^2:

$$n(n-1)p^2(p+q)^{n-2} = \sum_{m=0}^{n} m(m-1)C_n^m p^m q^{n-m}.$$

Учитывая, что $p+q=1$, имеем

$$n^2p^2 - np^2 = \sum_{m=0}^{n} m^2 C_n^m p^m q^{n-m} - \sum_{m=0}^{n} mC_n^m p^m q^{n-m}.$$

Используя равенства (3.26), (3.27) и (3.31), получаем выражение начального момента второго порядка:

$$\nu_2 = n^2p^2 - np^2 + np. \qquad (3.32)$$

Используя связь между центральными и начальными моментами, найдем выражение дисперсии и среднего квадратического отклонения для биномиального распределения:

$$\mu_2 = \sigma^2 = npq; \qquad (3.33)$$

$$\sigma = \sqrt{npq}. \qquad (3.34)$$

Показатели асимметрии и эксцесса для биномиального распределения имеют вид

$$A = (q-p)/\sqrt{npq}, \qquad (3.35)$$

$$E = (1-6pq)/(npq). \qquad (3.36)$$

Исследуем форму графиков биномиальных распределений: сначала при фиксированном n и меняющемся p, затем при фиксированном p и возрастающем n. На рис. 3.20 построены многоугольники биномиального распределения при $n=20$ и $p=0{,}1$; 0,3; 0,5; 0,7 и 0,9. Особенностью этих распределений является то, что вероятность $P_{m,n}$ сначала возрастает при увеличении m и достигает наибольшего значения при некотором вероятнейшем значении $m=m_0$, которое (как было доказано в § 2.10) можно определить из двойного неравенства $np-q \leqslant m_0 \leqslant np+p$.

Значение m_0 является модой биномиального распределения. Если имеются два вероятнейших значения, то распределение является *бимодальным*. Отметим, что для

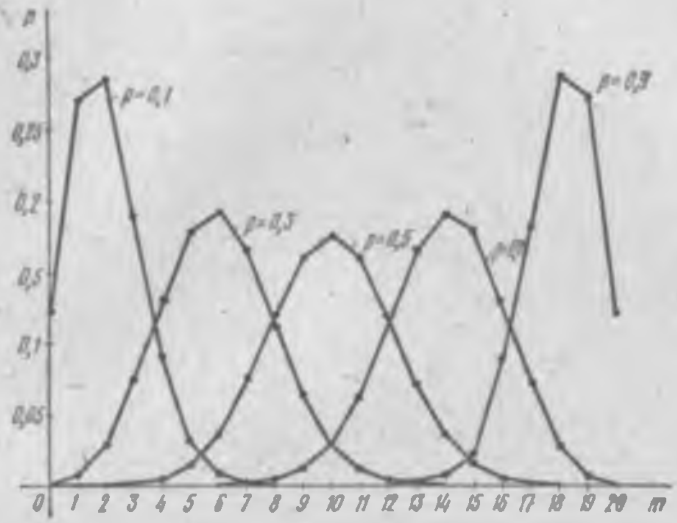

Рис. 3.20

любого биномиального распределения расстояние между математическим ожиданием и модой не превосходит единицы.

Если np — целое число, то математическое ожидание и мода совпадают. После достижения вероятнейшего значения m_0 вероятность $P_{m,n}$ начинает убывать. Распределение вообще асимметрично, за исключением случая, когда $p=0{,}5$. При $p<0{,}5$ асимметрия положительна. При $p>0{,}5$ асимметрия отрицательна. При увеличении числа испытаний n форма многоугольника распределения приближается к симметричной и при крайних значениях p. Это следует и из формул (3.35) и (3.36) (с ростом n

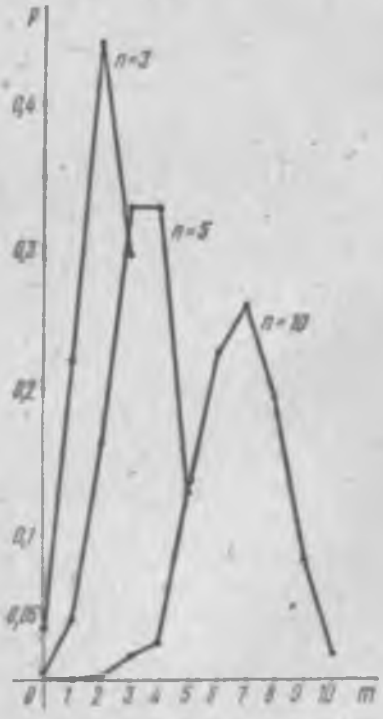

Рис. 3.21

асимметрия и эксцесс стремятся к нулю). Практически график биномиального распределения можно считать симметричным при $np \geqslant 4$. На рис. 3.21 изображены многоугольники распределения при $p=2/3$ и $n=3,5,10$. При возрастании n график биномиального распределения сдвигается вправо и становится более плоским.

§ 3.6. Локальная теорема Муавра—Лапласа

При большом числе испытаний n биномиальное распределение весьма близко к нормальному. Обоснование и доказательство этого положения содержатся в локальной теореме Муавра-Лапласа. Теорема для $p=1/2$ была доказана Муавром в 1730 г., а затем в 1783 г. обобщена Лапласом на случай произвольного p, отличного от нуля и единицы.

Локальная теорема Муавра—Лапласа. *Если вероятность p наступления события A в n независимых испытаниях постоянна и отлична от нуля и единицы, то при условии, что число испытаний достаточно велико, вероятность того, что в этих испытаниях событие A наступит равно m раз,*

$$P_{m,n} \approx \frac{1}{\sqrt{npq}} f(t), \qquad (3.37)$$

где $f(t) = \left(1/\sqrt{2\pi}\right) e^{-t^2/2}$; $t = (m-np)/\sqrt{npq}$.

Таким образом, из локальной теоремы Муавра—Лапласа следует, что распределение нормированной случайной величины $(m-np)/\sqrt{npq}$ стремится к нормальному при $n \to \infty$.

Локальная теорема Муавра—Лапласа относится к предельным теоремам теории вероятностей (см. гл. 4). Асимптотическая формула этой теоремы позволяет производить расчеты вероятности появлений событий при большом числе испытаний, не прибегая к формуле Бернулли.

Пример 3.11. Вероятность того, что станок-автомат производит годную деталь, равна 8/9. За смену изготовлено 280 деталей. Определить вероятность того, что среди них 20 бракованных.

Решение. Согласно условию задачи, $n=280$; $m=20$; $p=1/9$; $q=8/9$. По формуле (3.37) находим

$$P_{20,280} \approx \frac{1}{\sqrt{280 \cdot 1/9 \cdot 8/9}} f(t) = \frac{f(t)}{5,2588}.$$

Вычислим нормированное отклонение:

$$t = \frac{m - np}{\sqrt{npq}} = \frac{20 - 280 \cdot 1/9}{\sqrt{280 \cdot 1/9 \cdot 8/9}} = -2,11.$$

По табл. 1 приложений находим $f(-2,11) = 0,0431$. Искомая вероятность

$$P_{20,280} = 0,0431/5,2588 = 0,0082.$$

Вычисления, приведенные по формуле Бернулли, дадут такой же результат.

Пример 3.12. Используя условия примера 2.9, по формуле (3.37) вычислить вероятность того, что из пяти наудачу отобранных в течение смены деталей у одной размеры диаметра не соответствуют заданному допуску.

Р е ш е н и е. Имеем:

$$P_{1,5} \approx \frac{1}{\sqrt{5 \cdot 0,07 \cdot 0,93}} f(t) = \frac{f(t)}{0,5705};$$

$$t = \frac{m - np}{\sqrt{npq}} = \frac{1 - 5 \cdot 0,07}{\sqrt{5 \cdot 0,7 \cdot 0,93}} = \frac{0,65}{0,57} = 1,14.$$

По табл. 1 приложений находим $f(1,14) = 0,2083$. Искомая вероятность $P_{1,5} \approx 0,3651$.

Если вычисления проводить по формуле Бернулли, то приходим к другому результату, а именно: $P_{1,5} = 0,2618$ (см. пример 2.9 гл. 2). Такое расхождение объясняется тем, что формула локальной теоремы Муавра — Лапласа дает достаточно хорошие приближения лишь при больших значениях n ($n \geqslant 10$).

§ 3.7. Распределение Пуассона, или закон распределения редких явлений

Биномиальное распределение приближается к нормальному тем лучше, чем больше n и чем меньше p или q. Однако это не имеет места, если наряду с увеличением n одна из величин, p или q, стремится к нулю. В этом случае биномиальное распределение в пределе дает распределение Пуассона.

Преобразуем выражение для вероятности $P_{m,n}$ в биномиальном распределении, предварительно положив, что при увеличении n сохраняется равенство $a = np$, т. е. $p = a/n$. Имеем

$$P_{m,n} = \frac{n(n-1)(n-2)\ldots(n-m+1)(n-m)\ldots 2 \cdot 1}{m!\,(n-m)!} p^m q^{n-m} =$$

$$= \frac{1}{m!} n(n-1)(n-2)\ldots(n-m+1)\left(\frac{a}{n}\right)^m \left(1 - \frac{a}{n}\right)^{n-m} =$$

$$= \frac{1}{m!} \frac{n}{n} \frac{n-1}{n} \ldots \frac{n-m+1}{n} a^m \left(1 - \frac{a}{n}\right)^n \left(1 - \frac{a}{n}\right)^{-m}.$$

Вычислим пределы сомножителей:

$$\lim_{n\to\infty}\frac{n}{n}=1;\ \lim_{n\to\infty}\frac{n-1}{n}=\lim_{n\to\infty}\left(1-\frac{1}{n}\right)=1;$$

$$\lim_{n\to\infty}\frac{n-2}{n}=\lim_{n\to\infty}\left(1-\frac{2}{n}\right)=1;\ \lim_{n\to\infty}\frac{n-m+1}{n}=$$

$$=\lim_{n\to\infty}\left(1-\frac{m-1}{n}\right)=1;\ \lim_{n\to\infty}\left(1-\frac{a}{n}\right)^{-m}=1.$$

Для нахождения предела выражения $(1-a/n)^n$ обозначим $-a/n=1/x$, откуда $n=-ax$. Далее имеем

$$\lim_{n\to\infty}\left(1-\frac{a}{n}\right)^n=\lim_{x\to\infty}\left(1+\frac{1}{x}\right)^{-ax}=\lim_{x\to\infty}\left[\left(1+\frac{1}{x}\right)^x\right]^{-a}=e^{-a},$$

так как $\lim\limits_{x\to\infty}(1+1/x)^x=e=2{,}71828...$

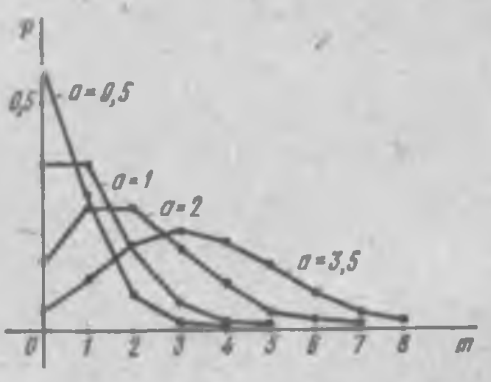

Рис. 3.22

Таким образом, при $n\to\infty$ и p или $q\to\infty$ получаем

$$P_m=a^m e^{-a}/m!. \qquad (3.38)$$

Это и есть формула распределения Пуассона. Распределение Пуассона описывает число событий m, происходящих за одинаковые промежутки времени при условии, что события происходят независимо друг от друга с постоянной средней интенсивностью. При этом число испытаний n велико, а вероятность появления события в каждом испытании p мала. Поэтому распределение Пуассона называется еще *законом распределения редких явлений*. Параметром распределения Пуассона является величина a, характеризующая интенсивность появления событий в n испытаниях. На рис. 3.22 приведе-

ны многоугольники распределения Пуассона, соответствующие различным значениям параметра a.

Пуассоновским распределением хорошо описывается распределение α-частиц, испускаемых радиоактивным источником за определенный промежуток времени, число требований на выплату страховых сумм за год, число вызовов, поступивших на телефонную станцию за определенное время суток, число отказов элементов при испытании на надежность сложных радиоэлектронных устройств, число бракованных изделий в выборках при взятии проб из партий изготовляемых на заводе изделий и т. д.

Распределение Пуассона может быть задано в виде ряда распределения:

$X=m$	0	1	2	...	m	...
P_m	e^{-a}	$\frac{a}{1!}e^{-a}$	$\frac{a}{2!}e^{-a}$	...	$\frac{a^m}{m!}e^{-a}$	...

Возможные значения случайной величины образуют бесконечную последовательность целых чисел 0, 1, 2, ...,

Можно убедиться, что при этом соблюдается необходимое требование для ряда распределения, сводящееся к тому, чтобы сумма всех вероятностей P_m была равна единице: $\sum_{m=0}^{\infty} P_m = e^{-a} e^{a} = 1$.

Найдем математическое ожидание и дисперсию случайной величины, распределенной по закону Пуассона:

$$M(x) = \sum_{m=0}^{\infty} mP_m = \sum_{m=0}^{\infty} m \frac{a^m}{m!} e^{-a} = a = np;$$

$$\sigma_x^2 = M(x^2) - [M(x)]^2 = a^2 + a - a^2 = a = np.$$

Отличительной особенностью этого распределения является равенство математического ожидания и дисперсии. Показатели асимметрии и эксцесса для распределения Пуассона имеют соответственно вид $A = 1/\sqrt{a}$; $E = 1/a$.

§ 3.8. Показательное распределение

Непрерывную случайную величину, плотность вероятности которой определяется выражением

$$f(x)=\begin{cases}\lambda e^{-\lambda x} & \text{при } x\geqslant 0,\\ 0 & \text{при } x<0,\end{cases}\quad \lambda>0, \tag{3.39}$$

называют *случайной величиной, имеющей показательное,* или *экспоненциальное, распределение.*

Это распределение часто наблюдается при изучении сроков службы различных устройств, как механических, так и электронных, времени безотказной работы отдельных элементов, части системы и системы в целом, т. е. при рассмотрении случайных промежутков времени между появлениями двух последовательных редких событий.

Плотность показательного распределения определяется параметром λ, который называют *интенсивностью отказов.* Этот термин связан с конкретной областью приложения распределения — теорией надежности*. Обратной величиной интенсивности отказов является наработка на отказ $1/\lambda$.

Выражение интегральной функции показательного распределения можно найти, используя свойство дифференциальной функции:

$$F(x)=\begin{cases}\int\limits_{-\infty}^{x} f(x)\,dx=\int\limits_{0}^{x}\lambda e^{-\lambda x}\,dx=1-e^{-\lambda x} & \text{при } x\geqslant 0,\\ 0 \text{ при } x>0.\end{cases} \tag{3.40}$$

Графики дифференциальной и интегральной функций распределения изображены соответственно на рис. 3.23, 3.24. Найдем выражения математического ожидания, дисперсии и среднего квадратического отклонения случайной величины с показательным распределением:

$$M(x)=\int\limits_{-\infty}^{\infty} xf(x)\,dx=\int\limits_{0}^{\infty}x\lambda e^{-\lambda x}\,dx=1/\lambda;$$

$$M(x^2)=\int\limits_{0}^{\infty}x^2\lambda\, e^{-\lambda x}\,dx=2/\lambda^2;$$

$$D(x)=M(x^2)-[M(x)]^2=2/\lambda^2-1/\lambda^2=1/\lambda^2;\ \sigma_x=1/\lambda.$$

* Под интенсивностью отказов в теории надежности понимают вероятность отказа невосстанавливаемого элемента в единицу времени после данного момента времени при условии, что до этого момента отказа не было.

Таким образом, для этого распределения характерно, что среднее квадратическое отклонение численно равно математическому ожиданию. Нетрудно убедиться, что при любом значении параметра λ коэффициенты вариации, асимметрии и эксцесса — постоянные величины: $V=1$; $A=2$; $E=9$.

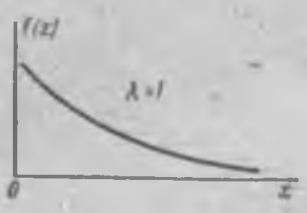

Рис. 3.23

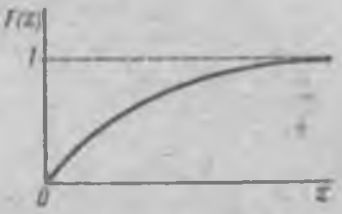

Рис. 3.24

§ 3.9. Понятие о системе случайных величин

На практике чаще всего сталкиваются с задачами, в которых результат опыта описывается двумя (и более) случайными величинами. В этом случае говорят о *двумерной* или *n-мерной* случайной величине, или о системе случайных величин. Например, прочностные свойства металла выражаются следующими показателями: твердостью, прочностью на разрыв, пределом текучести.

Система нескольких случайных величин X, Y, ..., Z обозначается так: $[X, Y, ..., Z]$. Каждая из величин, входящих в систему, называется ее *случайной составляющей*. Свойства системы нескольких случайных величин определяются не только свойствами отдельных ее составляющих, но и взаимными зависимостями между ними.

Систему двух случайных величин $[X, Y]$ геометрически можно представить как случайную точку на плоскости xOy с координатами X и Y. При геометрической интерпретации системы трех случайных величин используют трехмерное пространство. Систему n случайных величин рассматривают как случайную точку в n-мерном пространстве.

Ограничимся рассмотрением свойств системы двух случайных величин. Как было показано в § 3.1, распределение одномерной случайной величины задается либо в виде таблицы, либо в виде интегральной и дифференциальной функции в зависимости от типа случайной

величины. Аналогично может быть задано распределение системы случайных величин.

Если случайные величины X и Y дискретны с конечным числом возможных значений, то система $[X, Y]$ может быть задана в виде следующей таблицы:

Y \ X	x_1	x_2	...	x_i	...	x_n
y_1	p_{11}	p_{21}	...	p_{i1}	...	p_{n1}
y_2	p_{12}	p_{22}	...	p_{i2}	...	p_{n2}
. . .	. . .	. . .	...	. . .	...	. . .
y_j	p_{1j}	p_{2j}	...	p_{ij}	...	p_{nj}
. . .	. . .	. . .	...	. . .	...	. . .
y_m	p_{1m}	p_{2m}	...	p_{im}	...	p_{nm}

Здесь (x_i, y_i) — возможные пары значений случайных величин; $i=1, 2, ..., n$, $j=1, 2, ..., m$. Внутри таблицы распределения системы $[X, Y]$ указаны вероятности p_{ij} того, что случайная величина X примет значение x_i и одновременно с этим случайная величина Y примет значение y_j, т. е. $p_{ij}=P(X=x_i;\ Y=y_j)$.

Нетрудно убедиться, что имеет место равенство

$$\sum_{i,j} p_{ij} = \sum_{i,j} P\left(X=x_i;\ Y=y_j\right) = 1.$$

Наиболее универсальной формой задания распределения системы двух случайных величин является функция распределения, пригодная как для дискретных, так и для непрерывных случайных величин.

Функцией распределения системы двух случайных величин называется вероятность совместного выполнения неравенств $X<x$ и $Y<y$, т. е. $F(x,\ y)=P(X<x,\ Y<y)$. Если использовать геометрическую интерпретацию системы двух случайных величин в виде случайной

точки на плоскости xOy, то функция распределения $F(x, y)$ есть не что иное, как вероятность попадания случайной точки $(X; Y)$ в бесконечный квадрант с вершиной в точке $(x; y)$, лежащий левее и ниже ее (рис. 3.25). При такой интерпретации системы случайных величин $[X, Y]$ функция распределения случайной величины X представляет собой вероятность попадания случайной точки X в полуплоскость, ограниченную справа абсциссой x (рис. 3.26): $F_1(x)=P\{X<x\}$.

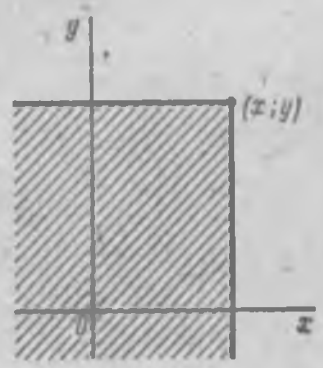

Рис. 3.25

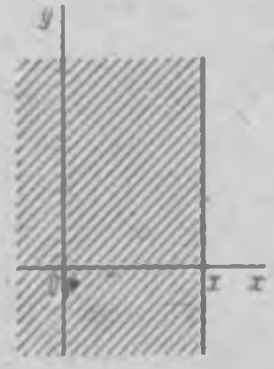

Рис. 3.26

Функция распределения случайной величины Y есть вероятность попадания случайной точки Y в полуплоскость, ограниченную ординатой y (рис. 3.27: $F_2(y)=P(Y<y)$.

Используя геометрическую интерпретацию функции распределения системы двух случайных величин $[X, Y]$, можно легко убедиться в следующих ее свойствах.

1°. *Функция распределения $F(x, y)$ является неубывающей функцией обоих своих аргументов, т. е.*

$$F(x_2, y) \geqslant F(x_1, y) \text{ при } x_2 > x_1;$$
$$F(x, y_2) \geqslant F(x, y_1) \text{ при } y_2 > y_1.$$

2°. *Если один из аргументов или оба равны минус-бесконечности, то функция распределения равна нулю, т. е.*

$$F(x, -\infty) = F(-\infty, y) = F(-\infty; -\infty) = 0;$$

3°. *Если один из аргументов равен плюс-бесконечности, то функция распределения системы переходит в*

функцию распределения случайной величины, соответствующей другому аргументу, т. е.

$$F(x; +\infty) = F_1(x);\ F(+\infty; y) = F_2(y).$$

4°. *Если оба аргумента равны плюс-бесконечности функция распределения системы равна единице, т. е.*

$$F(+\infty; +\infty) = 1.$$

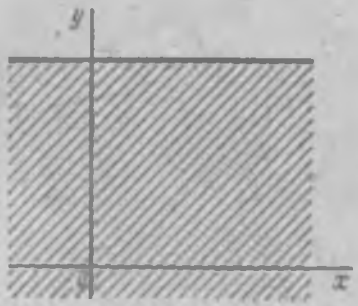

Рис. 3.27

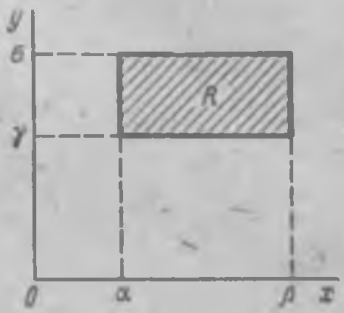

Рис. 3.28

Вероятность попадания случайной точки (X, Y) в прямоугольник R со сторонами, параллельными координатным осям, определяется по формуле

$$P(\alpha \leqslant X < \beta, \gamma \leqslant Y < \delta) = F(\beta, \delta) - F(\alpha, \delta) - \\ - F(\beta, \gamma) + F(\alpha, \gamma). \qquad (3.41)$$

Для вывода этой формулы следует использовать рис. 3.28.

Непрерывную двумерную случайную величину можно задавать с помощью дифференциальной функции или плотности распределения. Пусть система $[X, Y]$ состоит из двух непрерывных случайных величин X и Y. Применяя формулу (3.41), получаем вероятность попадания случайной точки $(X; Y)$ в элементарный прямоугольник со сторонами Δx и Δy, примыкающий к точке с координатами $(x; y)$ (рис. 3.29):

$$P(x<X<x+\Delta x,\ y<Y<y+\Delta y) = F(x+\Delta x, y+\Delta y) - \\ - F(x, y+\Delta y) - F(x+\Delta x, y) + F(x, y).$$

Разделим обе части этого равенства на значение площади этого прямоугольника и перейдем к пределу при $\Delta x \to 0$ и $\Delta y \to 0$. В пределе получаем вторую смешанную

частную производную от функции распределения $F(x, y)$ при условии, что $F(x, y)$ не только непрерывна, но и дважды дифференцируема:

$$\lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \frac{P(x < X < x + \Delta x,\ y < Y < y + \Delta y)}{\Delta x \Delta y} =$$

$$= \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \frac{F(x+\Delta x, y+\Delta y) - F(x, y+\Delta y) - F(x+\Delta x, y) + F(x, y)}{\Delta x \Delta y} =$$

$$= F''(x, y).$$

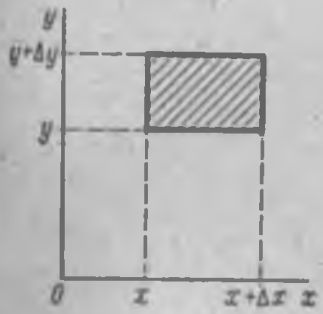

Рис. 3.29

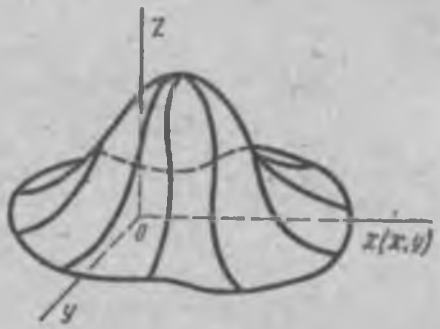

Рис. 3.30

Введем обозначение

$$f(x, y) = F''(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}.$$

Таким образом, *плотностью* или *дифференциальной функцией распределения системы двух случайных величин* $[X, Y]$ называют вторую смешанную частную производную от интегральной функции $F(x, y)$. Геометрически эту функцию можно истолковать как поверхность, поэтому ее называют *поверхностью распределения* (рис. 3.30).

При рассмотрении плотности распределения $f(x)$ одномерной случайной величины было введено понятие «элемента вероятности» $f(x)d(x)$. Для системы двух случайных величин элементом вероятности является выражение $f(x, y)\,dx\,dy$. Геометрически это есть не что иное как вероятность попадания случайной точки $(X; Y)$ в элементарный прямоугольник со сторонами dx и dy, прилегающий к точке $(x; y)$. Количественно эта вероятность равна объему элементарного параллелепипеда, ограни-

ченного сверху поверхностью $f(x, y)$ и опирающегося на элементарный прямоугольник $dxdy$ (рис. 3.31). По аналогии с формулой (3.31), используя введенное понятие элемента вероятности, выразим интегральную функцию системы двух случайных величин через дифференциальную:

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x, y)\, dx\, dy. \qquad (3.42)$$

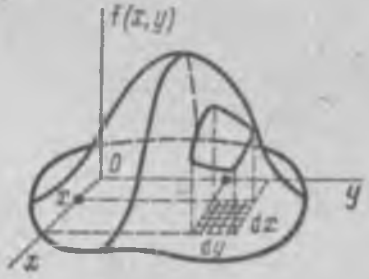

Рис. 3.31

Приведем некоторые свойства плотности распределения системы, в которых можно легко убедиться, используя приведенные выше выражения и геометрическую интерпретацию системы.

1°. *Плотность распределения системы есть функция неотрицательная:*

$$f(x, y) \geqslant 0.$$

2°. *Двойной интеграл в бесконечных пределах от плотности распределения системы равен единице:*

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1.$$

Покажем, что, зная закон распределения системы двух случайных величин, можно определить закон распределения составляющих ее случайных величин. Действительно, используя (3.42), запишем:

$$F_1(x) = F(x, \infty) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy.$$

Дифференцируя по x, получаем выражение плотности распределения случайной величины X:

$$f_1(x) = F_1'(x) = \int_{-\infty}^{\infty} f(x, y)\, dy. \qquad (3.43)$$

Аналогично находим выражения интегральной и дифференциальной функции распределения случайной величины Y:

$$F_2(y) = F(\infty, y) = \int_{-\infty}^{y} \int_{-\infty}^{\infty} f(x, y)\,dx\,dy;$$

$$f_2(y) = F_2'(y) = \int_{-\infty}^{\infty} f(x, y)\,dx. \qquad (3.44)$$

Таким образом, *плотность распределения* одной из величин, входящих в систему, получается путем интегрирования выражения плотности распределения системы в бесконечных пределах по аргументу, соответствующему другой случайной величине.

Решение обратной задачи — по известным законам распределения составляющих величин восстановить закон распределения системы — в общем случае невозможно, так как неизвестен характер взаимосвязи этих случайных величин. Поэтому вводятся так называемые условные законы распределения.

Условным законом распределения одной случайной величины, входящей в систему, называется закон, найденный при условии, что другая случайная величина, входящая в систему, приняла определенное значение.

Условный закон распределения задается как функцией распределения, так и плотностью распределения. Если рассматривается распределение случайной величины X при условии, что другая случайная величина Y приняла определенное значение, то условная функция распределения обозначается $F(x|y)$, а плотность — $f(x|y)$. Зная распределение одной из величин, входящей в систему, и условный закон распределения второй, можно найти распределение системы. По аналогии с теоремой умножения вероятностей (см. § 2.6) получаем:

$$f(x, y) = f_1(x)\,f(y|x); \qquad (3.45)$$

$$f(x, y) = f_2(y)\,f(x|y). \qquad (3.46)$$

Равенства (3.45), (3.46) выражают **теорему умножения законов распределения:** *плотность распределения системы двух случайных величин равна плотности распределения одной из величин, входящих в систему, умноженной на условную плотность распределения другой, вычисленную при условии, что первая величина приняла определенное значение.*

Используя равенства (3.43)—(3.46), запишем выражения условных плотностей распределения в таком виде:

$$f(y|x)=\frac{f(x,y)}{f_1(x)}=f(x,y)\Big/\int_{-\infty}^{\infty}f(x,y)\,dx; \qquad (3.47)$$

$$f(x|y)=\frac{f(x,y)}{f_2(y)}=f(x,y)\Big/\int_{-\infty}^{\infty}f(x,y)\,dx. \qquad (3.48)$$

Условные плотности распределения имеют те же свойства, что и безусловные плотности распределения.

Остановимся еще раз на уже введенном в § 3.1 понятии зависимых и независимых случайных величин. Случайная величина X называется независимой от случайной величины Y, если ее распределение вероятностей не зависит от того, какое значение приняла величина Y. Условие независимости X от Y может быть записано в виде

$$F(x|y)=F_1(x);\ f(x|y)=f_1(x) \qquad (3.49)$$

при любом y.

Если случайная величина X зависит от Y, то

$$F(x|y)\neq F_1(x);\ f(x|y)\neq f_1(x).$$

Легко показать используя равенства (3.45)—(3.46), что зависимость или независимость случайных величин всегда взаимна. Действительно,

$$f_1(x)f(y|x)=f_2(y)f(x|y).$$

Если выполняется условие независимости X от Y (3.49), то получаем

$$f(y|x)=f_2(y).$$

Для непрерывных случайных величин X и Y можно сформулировать следующий признак их независимости: *для того чтобы непрерывные случайные величины X и Y были независимыми, необходимо и достаточно, чтобы плотность распределения системы $[X, Y]$ была равна произведению плотностей распределения отдельных величин, входящих в систему:*

$$f(x,y)=f_1(x)f_2(y).$$

Рассмотрим числовые характеристики системы двух случайных величин. Введем понятие начальных и центральных моментов. *Начальным моментом порядка $k+s$* системы $[X, Y]$ называют математическое ожидание произведения X^k на Y^s:

$$\nu_{k,s} = M[X^k Y^s].$$

Наиболее употребительными являются начальные моменты первого порядка:

$$\nu_{1,0} = M[X^1 Y^0] = M(x), \quad \nu_{0,1} = M[X^0 Y^1] = M(y),$$

которые являются математическими ожиданиями случайных величин X и Y. Совокупность математических ожиданий представляет собой характеристику положения системы: $M(x)$ и $M(y)$ — координаты средней точки на плоскости, вокруг которой происходит рассеяние системы.

Центральным моментом порядка $k+s$ называют математическое ожидание произведения k-й и s-й степеней соответствующих центрированных величин:

$$\mu_{k,s} = M[\mathring{X}^k \mathring{Y}^s] = M\{[X - M(x)]^k [Y - M(y)]^s\}.$$

Наиболее употребительны центральные моменты второго порядка

$$\mu_{2,0} = M[\mathring{X}^2 \mathring{Y}^0] = M[X - M(x)]^2 = \sigma_x^2,$$

$$\mu_{0,2} = M[\mathring{X}^0 \mathring{Y}^2] = M[Y - M(y)]^2 = \sigma_y^2,$$

которые совпадают с дисперсиями случайных величин X и Y. Они характеризуют рассеяние системы $[X, Y]$ в направлении осей Ox и Oy.

Важными характеристиками являются условные математические ожидания и условные дисперсии. *Условным математическим ожиданием дискретной случайной величины X* при $Y=y$ называют сумму произведений возможных значений X на их условные вероятности:

$$M(X \mid Y = y) = \sum_{i=1}^{n} x_i P(x_i \mid y).$$

Для непрерывных случайных величин

$$M(X \mid Y = y) = \int_{-\infty}^{\infty} x f(x \mid y)\, dx.$$

Аналогично можно написать выражения для $M(Y|X=x)$:

$$M(Y|X=x)=\sum_{j=1}^{m} y_j P(y_j|x),$$

$$M(Y|X=x)=\int_{-\infty}^{\infty} yf(y|x)\,dy.$$

В случае дискретных случайных величин условная дисперсия

$$\sigma^2(X|Y=y)=\sum_{i=1}^{n}\{x_i-M(X|Y=y)\}^2 P(x_i|y).$$

Для непрерывных случайных величин

$$\sigma^2(X|Y=y)=\int_{-\infty}^{\infty}\{x-M(X|Y=y)\}^2 f(x|y)\,dx.$$

Аналогично записываются выражения условной дисперсии $\sigma^2(Y|X=x)$:

$$\sigma^2(Y|X=x)=\sum_{j=1}^{m}\{y_j-M(Y|X=x)\}^2 P(y_j|x),$$

$$\sigma^2(Y|X=x)=\int_{-\infty}^{\infty}\{y-M(Y|X=x)\}^2 f(y|x)\,dy.$$

Особая роль при изучении системы двух случайных величин принадлежит второму смешанному центральному моменту $\mu_{11}=M[\mathring{X}\mathring{Y}]$, который представляет собой математическое ожидание произведения центрированных случайных величин K_{xy} (*корреляционный момент или момент связи* X и Y):

$$K_{xy}=M[\mathring{X}\mathring{Y}]=M\{[X-M(x)][Y-M(y)]\}.$$

Корреляционный момент характеризует рассеяние и зависимость случайных величин. Если, например, одна из составляющих величин незначительно отклоняется от своего математического ожидания, то корреляционный момент мал, как бы тесно ни были связаны X и Y.

Чтобы избежать влияния рассеяния случайных величин на корреляционный момент и получить показатель

характеристики связи между величинами X и Y, переходят от момента K_{xy} к безразмерной характеристике

$$\rho_{xy} = K_{xy}/(\sigma_x \sigma_y), \qquad (3.50)$$

которую называют *коэффициентом корреляции* случайных величин X и Y*.

Для независимых случайных величин коэффициент корреляции и корреляционный момент равны нулю. Покажем это для непрерывных случайных величин. Для независимых случайных величин $f(x, y) = f_1(x) f_2(y)$, где $f_1(x)$ и $f_2(y)$ — соответствующие плотности распределения вероятностей величин X и Y. Тогда

$$K_{xy} = \int_{-\infty}^{\infty}\int [X - M(x)][Y - M(y)] f(x, y)\, dx\, dy =$$

$$= \int_{-\infty}^{\infty} [X - M(x)] f_1(x)\, dx \int_{-\infty}^{\infty} [Y - M(y)] f_2(y)\, dy =$$

$$= M[\mathring{X}]\, M[\mathring{Y}] = 0,$$

следовательно, и $\rho_{xy} = 0$. □

Обратное утверждение — если коэффициент корреляции между двумя случайными величинами равен нулю, то эти случайные величины независимы — не всегда справедливо. Это значит, что может существовать система зависимых случайных величин, коэффициент корреляции которых равен нулю. Поэтому вводится понятие коррелированности.

Две случайные величины называются *коррелированными*, если их коэффициент корреляции отличен от нуля. Величины X и Y некоррелированы, если их коэффициент корреляции равен нулю.

Таким образом, если случайные величины X и Y независимы, то они и некоррелированы. Но если они некоррелированы, то это не является достаточным условием для их независимости. Лишь при нормальном распределении величин некоррелированность и независимость тождественны.

---

* Так как выражение коэффициента корреляции симметрично относительно X и Y, т. е. $\rho_{xy} = \rho_{yx}$, то в дальнейшем индекс при ρ опускаем.

§ 3.10. Двумерное нормальное распределение

Рассмотрим нормальное распределение системы двух случайных величин. Так как система двух случайных величин изображается случайной точкой на плоскости, то нормальное распределение системы двух величин на-

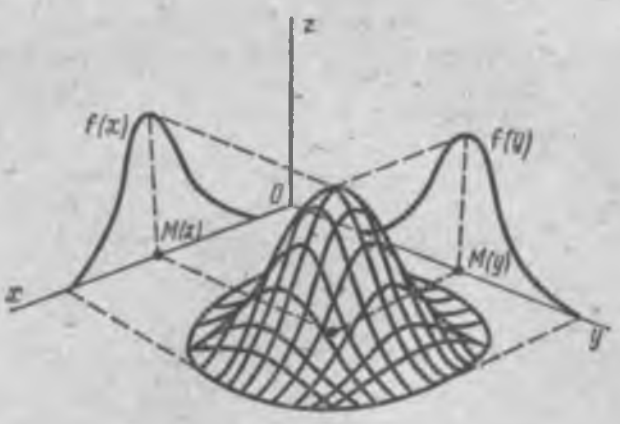

Рис. 3.32

зывают нормальным распределением на плоскости. Плотность двумерного нормального распределения

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times$$

$$\times e^{-\frac{\left\{\frac{[x-M(x)]^2}{\sigma_x^2} + \frac{[y-M(y)]^2}{\sigma_y^2} - 2\rho\frac{[x-M(x)][y-M(y)]}{\sigma_x\sigma_y}\right\}}{2(1-\rho^2)}}. \quad (3.51)$$

Из выражения плотности (3.51) следует, что двумерное нормальное распределение определяется параметрами $M(x)$, $M(y)$, σ_x, σ_y и коэффициентом корреляции ρ, определяемым по формуле (3.50). Плотность двумерного нормального распределения изображается в системе координат поверхностью, которая носит название *поверхности нормального распределения* (рис. 3.32). В сечении нормальной поверхности распределения плоскостями, параллельными координатной плоскости xOz, получаются кривые распределения случайной величины X, соответствующие определенным значениям y (рис. 3.33). Аналогично, в сечении нормальной поверхности распределения плоскостями, параллельными координатной

плоскости yOz, получаются кривые распределения случайной величины Y, соответствующие определенным значениям x. Эти кривые являются графическими изображениями условных законов распределений соответственно случайных величин X и Y при фиксированных значениях y и x.

Математические ожидания этих условных законов распределения являются условными математическими

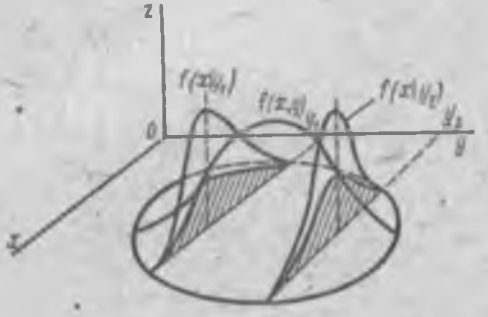

Рис. 3.33

ожиданиями. Если спроектировать на плоскость xOy условные математические ожидания $M[X|Y=y]$ и соединить линией полученные точки, то образованная таким образом линия называется линией регрессии X по Y. Таким образом, линия регрессии X по Y — множество точек, соответствующих условным математическим ожиданиям $M[X|Y=y]$. Линия регрессии Y по X — множество точек, соответствующих условным математическим ожиданиям $M[Y|X=x]$.

Безусловные законы распределения составляющих систему случайных величин X и Y являются нормальными. На рис. 3.32 кривые распределения безусловных законов распределения спроектированы на плоскостях xOz и yOz.

Плотности безусловных законов распределения случайных величин X и Y, входящих в систему, соответственно равны:

$$f_1(x) = \frac{1}{\sqrt{2\pi}\,\sigma_x} e^{-[x-M(x)]^2/\left(2\sigma_x^2\right)}, \qquad (3.52)$$

$$f_2(y) = \frac{1}{\sqrt{2\pi}\,\sigma_y} e^{-[y-M(y)]^2/\left(2\sigma_y^2\right)}. \qquad (3.53)$$

Подставляя в выражения (3.47)—(3.48) формулы (3.51)—(3.53), получаем условные плотности распределения случайных величин Y и X:

$$f(y|x) = \frac{2}{\sqrt{2\pi}\,\sigma_y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}\left\{\frac{y-M(y)-\rho\frac{\sigma_y}{\sigma_x}[x-M(x)]}{\sigma_y\sqrt{1-\rho^2}}\right\}^2}, \tag{3.54}$$

$$f(x|y) = \frac{2}{\sqrt{2\pi}\,\sigma_x\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}\left\{\frac{x-M(x)-\rho\frac{\sigma_x}{\sigma_y}[y-M(y)]}{\sigma_x\sqrt{1-\rho^2}}\right\}^2}. \tag{3.55}$$

Таким образом, видим, что условные распределения случайных величин Y и X, если их совместное распределение нормально, также являются нормальными с центрами (условными математическими ожиданиями):

$$W(Y|X=x) = M(y) + \rho\frac{\sigma_y}{\sigma_x}[x - M(x)]; \tag{3.56}$$

$$M[X|Y=y] = M(x) + \rho\frac{\sigma_x}{\sigma_y}[y - M(y)]. \tag{3.57}$$

Условные плотности распределения:

$$f(y|x) = \frac{1}{\sqrt{2\pi}\,\sigma_{y|x}}\, e^{-\frac{1}{2}\left[\frac{y-M(y|x)}{\sigma_{y|x}}\right]^2},$$

$$f(x|y) = \frac{1}{\sqrt{2\pi}\,\sigma_{x|y}}\, e^{-\frac{1}{2}\left[\frac{x-M(x|y)}{\sigma_{x|y}}\right]^2},$$

где $\sigma_{y|x}$ и $\sigma_{x|y}$ — соответствующие условные средние квадратические отклонения:

$$\sigma_{y|x} = \sigma_y\sqrt{1-\rho^2}, \tag{3.58}$$

$$\sigma_{x|y} = \sigma_x\sqrt{1-\rho^2}. \tag{3.59}$$

Из выражений (3.56) и (3.57) следует, что линии регрессии в случае двумерного нормального распределения являются прямыми.

В плоскости xOy уравнение линии регрессии Y по X имеет вид

$$y - M(y) = \rho \frac{\sigma_y}{\sigma_x} [x - M(x)].$$

Уравнение линии регрессии X по Y таково:

$$x - M(x) = \rho \frac{\sigma_x}{\sigma_y} [y - M(y)].$$

Из выражений (3.58) и (3.59) следует, что рассеяние условных распределений составляющих величин Y и X при всех значениях другой переменной постоянно. Это

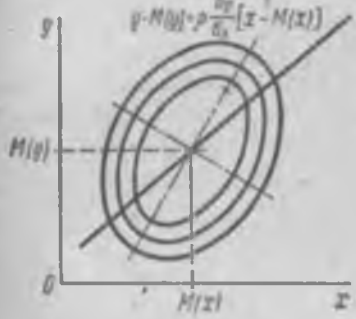

Рис. 3.34

Рис. 3.35

свойство называется *гомоскедастичностью*, т. е. *равноизменчивостью* условных нормальных распределений.

Пересекая поверхность распределения плоскостями $z=z_0=\text{const}$, параллельными координатной плоскости xOy, в проекции на этой плоскости получаем семейство концентрических эллипсов различных размеров с одинаковой ориентацией главных осей и с общим центром в точке с координатами $[M(x); M(y)]$ (рис. 3.34). Каждый эллипс этого семейства является множеством точек, где плотность распределения $f(x, y)$ постоянна. Поэтому их называют *эллипсами равной плотности* или *эллипсами рассеяния*.

Ориентация эллипсов рассеяния относительно координатных осей находится в прямой зависимости от величины коэффициента корреляции. При $\rho=\pm 1$ эллипсы вырождаются в прямую*. Если случайные величины X и Y

* В § 9.4 показано, что $-1 \leqslant \rho \leqslant 1$.

некоррелированы, т. е. $\rho_{xy}=0$, то главные оси (их также называют главными осями рассеяния) параллельны осям Ox и Oy (рис. 3.35). С увеличением корреляционной зависимости между величинами X и Y происходит все больший поворот главных осей рассеяния относительно координатных осей.

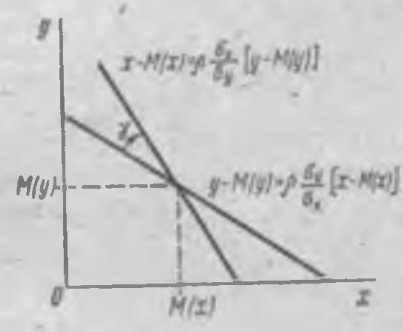

Рис. 3.36

Линии регрессии Y по X и X по Y проходят через точку с координатами $[M(x); M(y)]$, которая является точкой их пересечения и центром эллипсов рассеяния. Следует подчеркнуть, что главные оси рассеяния эллипсов не могут быть параллельны линиям регрессии. Угловые коэффициенты линий регресии, получившие названия *коэффициентов регрессии*, таковы:

$$\beta_{y|x}=\rho\frac{\sigma_y}{\sigma_x}; \quad (3.60) \qquad \beta_{x|y}=\rho\frac{\sigma_x}{\sigma_y}. \quad (3.61)$$

Произведение коэффициентов регрессии равно квадрату коэффициента корреляции: $\beta_{y|x}\beta_{x|y}=\rho^2$. Из выражений (3.60) и (3.61) следует, что коэффициенты регрессии и коэффициент корреляции имеют одинаковые знаки. Если обозначить через γ острый угол между линиями регрессии Y по X и X по Y (рис. 3.36), то можно показать, что имеет место соотношение

$$\operatorname{tg}\gamma=\frac{1-\rho^2}{\rho}\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}. \quad (3.62)$$

Из формулы (3.62) следует, что чем меньше коэффициент корреляции, тем больше угол между прямыми регрессии. При $\rho=0$, т. е. если корреляционная связь между X и Y отсутствует, имеем $\operatorname{tg}\gamma=\infty$. Следовательно, эти две прямые взаимно перпендикулярны. При $\rho=1$ прямые совпадают, тогда $\beta_{x|y}=1/\beta_{y|x}$.

В § 3.9 отмечалось, что в общем случае условие равенства коэффициента корреляции нулю недостаточно для независимости случайных величин. Покажем, что *если нормально распределенные случайные величины некор-*

релированы, то они и независимы. Действительно, при $\rho=0$ выражение плотности двумерного нормального распределения (3.50) имеет вид

$$f(x, y)=\frac{1}{\sqrt{2\pi}\,\sigma_x}e^{-[x-M(x)]^2/\left(2\sigma_x^2\right)}\frac{1}{\sqrt{2\pi}\,\sigma_y}e^{-[y-M(y)]^2/\left(2\sigma_y^2\right)},$$

т. е. $f(x, y)=f_1(x)f_2(y)$. Тогда в соответствии со сформулированным в § 3.9 признаком независимости можно утверждать, что X и Y — независимые случайные величины.

Глава 4

Предельные теоремы теории вероятностей

§ 4.1. Предварительные замечания

Теория вероятностей изучает закономерности, свойственные массовым случайным явлениям. Предельные теоремы теории вероятностей устанавливают зависимость между случайностью и необходимостью. Изучение закономерностей, проявляющихся в массовых случайных явлениях, позволяет научно предсказывать результаты будущих испытаний. Предельные теоремы теории вероятностей делятся на две группы, одна из которых получила название *закона больших чисел*, а другая — *центральной предельной теоремы*.

В настоящей главе рассмотрены следующие теоремы, относящиеся к закону больших чисел: лемма Маркова и неравенство Чебышева, теоремы Чебышева, Маркова, Бернулли и Пуассона.

Закон больших чисел состоит из нескольких теорем, в которых доказывается приближение средних характеристик при соблюдении определенных условий к некоторым постоянным значениям. Впервые доказал и сформулировал одну из теорем закона больших чисел швейцарский математик Яков Бернулли (1654—1705). Теорема, носящая его имя, одна из простейших форм закона больших чисел, касается альтернативной изменчивости признака и устанавливает закономерность частности появления события.

Выдающийся французский математик Симеон Дени Пуассон (1781—1840) в одной из своих работ обобщил эту теорему, распространив ее на случай, когда вероятность события в независимых испытаниях изменяется. Он же впервые ввел термин «закон больших чисел».

Великий русский математик П. Л. Чебышев (1821—1894) доказал, что закон больших чисел действует в яв-

лениях с любой вариацией и распространяется также на среднюю величину. Дальнейшее обобщение теорем закона больших чисел связано с именами А. А. Маркова, С. Н. Бернштейна, А. Я. Хинчина и А. Н. Колмогорова.

Рассмотрение теорем закона больших чисел начнем с леммы Маркова и неравенства Чебышева, с помощью которых значительно упрощаются доказательства теорем закона больших чисел.

§ 4.2. Лемма Маркова и неравенство Чебышева

Лемма Маркова. *Для любой положительной случайной величины X вероятность того, что она примет значение, не превосходящее некоторого положительного числа τ, больше, чем разность между единицей и отношением математического ожидания этой случайной величины к данному числу τ:*

$$P\{X \leqslant \tau\} > 1 - M(x)/\tau. \qquad (4.1)$$

Доказательство. Проведем доказательство для дискретной случайной величины способом, получившим название *метода урезания*. Пусть $x_1, x_2, \ldots, x_n$ — упорядоченная совокупность всех значений, принимаемых положительной случайной величиной X с соответствующими вероятностями $p_1, p_2, \ldots, p_n$, причем $\sum_{i=1}^{n} p_i = 1$. Не нарушая общности доказательства, можно допустить, что значения случайной величины X расположены в порядке убывания. Выберем некоторое произвольное число $\tau > 0$ и предположим что первые r значений совокупности больше $\tau (r \leqslant n)$. Так как по условию случайная величина X принимает только положительные значения, то можем записать неравенство

$$x_1 p_1 + x_2 p_2 + \ldots + x_r p_r \leqslant x_1 p_1 + x_2 p_2 + \ldots + x_r p_r + \ldots + x_n p_n.$$

По определению математического ожидания, $x_1 p_1 + x_2 p_2 + \ldots + x_r p_r + \ldots + x_n p_n = M(x)$, следовательно, $x_1 p_1 + x_2 p_2 + \ldots + x_r p_r \leqslant M(x)$. Заменяя в левой части последнего неравенства значения переменных $x_i (i = 1, 2, \ldots, r)$ числом τ, получаем следующее усиленное неравенство:

$$(p_1 + p_2 + \cdots + p_r)\tau \leqslant M(x), \text{ или}$$

$$p_1 + p_2 + \ldots + p_r \leqslant M(x)/\tau.$$

Левая часть этого неравенства выражает вероятность

того, что случайная величина принимает значение, большее τ, т. е. $P\{X>\tau\}$. Поэтому полученное неравенство можно записать в виде

$$P\{X>\tau\}\leqslant M(x)/\tau.$$

Вероятность противоположного события, а именно вероятность того, что случайная величина X может принимать значения не больше τ, определяется следующим неравенством:

$$P\{X\leqslant r\}>1-M(x)/\tau. \quad \square$$

Лемма Маркова справедлива для любого распределения положительной случайной величины.

Неравенство Чебышева. *Если случайная величина X имеет конечное математическое ожидание и дисперсию, то для любого положительного числа τ справедливо неравенство*

$$P\{|X-M(x)|\leqslant\tau\}>1-\sigma_x^2/\tau^2, \qquad (4.2)$$

т. е. вероятность того, что отклонение случайной величины X от своего математического ожидания по абсолютной величине не превзойдет τ, больше разности между единицей и отношением дисперсии этой случайной величины к квадрату τ.

Доказательство. Воспользуемся леммой Маркова. Рассмотрим случайную величину $Z=|X-M(x)|$, для некоторых значений которой выполняется неравенство $|X-M(x)|\leqslant\tau$. Так как случайная величина Z положительна, то неравенства $|X-M(x)\leqslant\tau|$ и $[X-M(x)]^2\leqslant\tau^2$ равносильны. Применив, лемму Маркова к случайной величине $Z^2=[X-M(x)]^2$, получим

$$P\{[X-M(x)]^2\leqslant\tau^2\}>1-M[X-M(x)]^2/\tau^2.$$

Числитель дроби в правой части неравенства, по определению, есть дисперсия случайной величины X, следовательно, неравенство можно записать в виде

$$P\{[X-M(x)]^2\leqslant\tau^2\}>1-\sigma_x^2/\tau^2.$$

Так как в силу равносильности соответствующих неравенств вероятности $P\{[X-M(x)]^2\leqslant\tau^2\}$ и $P\{|X-M(x)|\leqslant\tau\}$ равны, то

$$P\{|X-M(x)|\leqslant\tau\}>1-\sigma_x^2/\tau^2. \quad \square$$

Запишем теперь вероятность события $|X-M(x)|>\tau$, т. е. события, противоположного событию $|X-M(x)|\leqslant\tau$. Очевидно, что

$$P\{|X-M(x)|\leqslant\tau\}>\sigma_x^2/\tau^2. \qquad (4.3)$$

Неравенство Чебышева, как и неравенство Маркова, справедливо для любого распределения случайной величины X, однако неравенство Чебышева применимо как к положительным, так и к отрицательным случайным величинам. Неравенство (4.3) ограничивает сверху вероятность того, что случайная величина отклонится от своего математического ожидания на величину, большую τ. Из этого неравенства следует, что при уменьшении дисперсии верхняя граница вероятности также уменьшается и значения случайной величины с небольшой дисперсией сосредоточиваются около ее математического ожидания. Полагая $\tau=t\sigma$, запишем неравенства (4.2) и (4.3) в виде

$$P\{|X-M(x)|\leqslant t\sigma\}>1-1/t^2,$$
$$P\{|X-M(x)|>t\sigma\}\leqslant 1/t^2.$$

Задаваясь различными значениями t, вычислим верхние границы вероятностей того, что отклонение случайной величины выйдем за пределы $t\sigma$: при $t=1$ $P\{|X-M(x)|>\sigma\}\leqslant 1$; $t\neq 2$ $P\{|X-M(x)|>2\sigma\}\leqslant 1/4$; $t=3$ $P\{|X-M(x)|>3\sigma\}\leqslant 1/9$.

В заключение еще раз подчеркнем, что вероятности рассматриваемых событий не могут превысить этих числовых значений ни при каком распределении случайной величины.

§ 4.3. Теорема Чебышева

Теорема Чебышева. *При достаточно большом числе попарно независимых случайных величин X_1, X_2, ..., X_n с математическими ожиданиями $M(x_1)$, $M(x_2)$, ..., $M(x_n)$ и дисперсиями $\sigma_{x_1}^2$, $\sigma_{x_2}^2$, ..., $\sigma_{x_n}^2$ с вероятностью, близкой единице, можно утверждать, что разность между средней арифметической наблюдавшихся значений случайных величин X_1, X_2, ..., X_n и средней арифметической их математических ожиданий по абсолютной величине окажется меньше сколь угодно малого числа $\tau>0$ при условии,*

что дисперсия всех этих случайных величин не превосходит одного и того же постоянного числа B, т. е.

$$P\left\{\left|\sum_{i=1}^{n} X_i/n - \sum_{i=1}^{n} M(x_i)/n\right| \leqslant \tau\right\} > 1 - \eta, \qquad (4.4)$$

где η — *положительное число, близкое к нулю.*

Доказательство. Пусть X_1, X_2, ..., X_n—n попарно независимых случайных величин. Средняя арифметическая этих величин является, в свою очередь, тоже случайной величиной. Обозначим ее $\overline{X}$:

$$\overline{X} = \sum_{i=1}^{n} X_i/n.$$

Определим математическое ожидание и дисперсию $\overline{X}$:

$$M(\overline{x}) = M\left[\sum_{i=1}^{n} X_i/n\right] = \sum_{i=1}^{n} M(x_i)/n, \qquad (4.5)$$

$$\sigma_{\overline{x}}^2 = \sigma^2\left[\sum_{i=1}^{n} X_i/n\right] = \sum_{i=1}^{n} \sigma_{x_i}^2/n^2. \qquad (4.6)$$

Применяя к случайной величине $\overline{X}$ неравенство Чебышева, получаем

$$P\{|\overline{X} - M(\overline{x})| \leqslant \tau\} > 1 - \sigma_{\overline{x}}^2/\tau^2,$$

или

$$P\left\{\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} M(x_i)}{n}\right| \leqslant \tau\right\} > 1 - \frac{\sum_{i=1}^{n} \sigma_{x_i}^2}{n^2\tau^2}. \qquad (4.7)$$

Заменяя в правой части последнего неравенства дисперсии случайных величин величиной B, которую по условию теоремы не превосходит ни одна из дисперсий, получаем усиленное неравенство

$$P\left\{\left|\sum_{i=1}^{n} X_i/n - \sum_{i=1}^{n} M(x_i)/n\right| \leqslant \tau\right\} > 1 - \frac{B}{n\tau^2}. \qquad (4.8)$$

Для любого сколь угодно малого числа $\eta > 0$ можно найти такое число n, при котором выполняется неравен-

ство $B/(n\tau^2)<\eta$. Таким образом получаем неравенство, справедливость которого и требовалось доказать:

$$P\left\{\left|\frac{\sum_{i=1}^{n} X_i}{n}-\frac{\sum_{i=1}^{n} M(x_i)}{n}\right|\leqslant\tau\right\}>1-\eta. \quad \square$$

Переходя в фигурных скобках к противоположному событию, можно записать

$$P\left\{\left|\frac{\sum_{i=1}^{n} X_i}{n}-\frac{\sum_{i=1}^{n} M(x_i)}{n}\right|>\tau\right\}\leqslant\eta. \qquad (4.9)$$

Итак, теорема Чебышева показывает, что средняя арифметическая попарно независимых случайных величин обладает свойством устойчивости и при определенных условиях мало отличается от средней арифметической математических ожиданий этих величин.

Рассмотрим частный случай теоремы. Наблюдавшиеся при n независимых испытаниях значения случайной величины X с математическим ожиданием $M(x)$ и дисперсией σ_x^2 можно рассматривать как случайные величины X_1, X_2, ..., X_n с одинаковыми математическими ожиданиями и дисперсиями: $M(x_i)=M(x)$; $\sigma_{x_i}^2=\sigma_x^2$ ($i=1, 2, ..., n$). Средняя арифметическая этих случайных величин $\overline{X}=\sum_{i=1}^{n} X_i/n$. Используя свойства математического ожидания и дисперсии случайной величины, а также формулы (4.5) и (4.6), можно показать, что $M(\overline{x})=M(x)$ и $\sigma_{\overline{x}}^2=\sigma_x^2/n$. Тогда теорему Чебышева можно записать в следующем виде:

$$P\{|\overline{X}-M(x)|\leqslant\tau\}\geqslant 1-\sigma_x^2/(\tau^2 n).$$

Пример 4.1. На завод поступила партия деталей, рассортированных автоматом для селективной сборки по их длине. Детали, распределенные по размерам на группы, помещены в 64 ящика. При этом во всех ящиках находится одинаковое количество деталей. Из каждого ящика отобрано по 10 деталей и измерены их длины. При условии, что дисперсии размеров контролируемого параметра в каждом ящике не превышают 0,015 мм², определить, с какой вероятностью можно ожидать, что средняя арифметическая, вычисленная по результатам выборок, отклонится по абсолютной величине от средней длины деталей во всей партии не более чем на 0,3 мм.

Р е ш е н и е. По условию, $n=640$, $\tau=0{,}3$, $B=0{,}015$. Воспользовавшись неравенством (4.8), получим

$$P\left\{\left|\frac{\sum_{i=1}^{n} X_i}{n}-\frac{\sum_{i=1}^{n} M(x_i)}{n}\right|<0{,}3\right\}>0{,}9997.$$

§ 4.4. Теоремы Бернулли и Пуассона

Теорема Бернулли устанавливает связь между частостью появления события и его вероятностью. Она может быть выведена как следствие из теоремы Чебышева.

Теорема Бернулли. *При достаточно большом числе независимых испытаний n можно с вероятностью, близкой к единице, утверждать, что разность между частостью появления события A в этих испытаниях и его вероятностью в отдельном испытании по абсолютной величине окажется меньше сколь угодно малого числа τ, если вероятность наступления этого события в каждом испытании постоянна и равна p.*

Утверждение теоремы можно записать в виде следующего неравенства:

$$P\{|m/n-p|\leqslant\tau\}>1-\eta, \tag{4.10}$$

где τ и η — любые сколь угодно малые положительные числа.

Д о к а з а т е л ь с т в о. Рассмотрим в качестве случайных величин число появлений события A в n испытаниях. Пусть X_1 — число появлений события A в первом испытании; X_2 — число появлений события A во втором испытании; X_n — число появлений события A в n-м испытании. Так как событие A в каждом испытании может либо произойти, либо не произойти, то, следовательно, случайная величина может принимать лишь два значения: 1 и 0 с соответствующими вероятностями p и $q=1-p$.

Математическое ожидание и дисперсия каждой случайной величины соответственно равны:

$$M(x_i)=1\cdot p+0\cdot q=p, \tag{4.11}$$

$$\sigma^2_{x_i}=(1-p)^2\cdot p+(0-p)^2\cdot q=pq. \tag{4.12}$$

Очевидно, что

$$M(x_1) = M(x_2) = \ldots = M(x_n), \tag{4.13}$$

$$\sigma^2_{x_1} = \sigma^2_{x_2} = \ldots = \sigma^2_{x_n}. \tag{4.14}$$

Исходя из того, что в силу независимости испытаний обеспечивается взаимная независимость случайных величин X_1, X_2, ..., X_n и дисперсии всех величин ограничены (произведение pq имеет наибольшее значение, равное $1/4$)*, к рассматриваемым величинам можно применить теорему Чебышева. В результате, учитывая равенства (4.11) — (4.14) и тот факт, что средняя арифметическая величин X_1, X_2, ..., X_n есть не что иное, как частость появлений события A в n испытаниях, т. е. $\sum_{i=1}^{n} X_i/n = m/n$, получаем следующее неравенство:

$$P\{|m/n - p| \leqslant \tau\} > 1 - (pq)/(n\tau^2). \tag{4.15}$$

Для любого сколь угодно малого $\eta > 0$ всегда можно определить такое n, при котором выполняется неравенство $(pq)/(n\tau^2) < \eta$. Итак, доказана справедливость неравенства

$$P\{|m/n - p| \leqslant \tau\} > 1 - \eta. \quad \square$$

При решении практических задач иногда бывает необходимо оценить вероятность наибольшего отклонения частоты появлений события от ее ожидаемого значения. Случайной величиной в этом случае является число появлений события A в n независимых испытаниях. Имеем:

$$m = X_1 + X_2 + \ldots + X_n;$$

$$M(m) = M\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} M(x_i) = \sum_{i=1}^{n} p = np;$$

$$\sigma^2_m = \sigma^2\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \sigma^2[X_i] = \sum_{i=1}^{n} pq = npq.$$

Используя неравенство Чебышева, в этом случае получаем

$$P\{|m - np| \leqslant \tau\} > 1 - (npq)/\tau^2. \tag{4.16}$$

* Можно доказать, что произведение двух сомножителей, сумма которых — величина постоянная, имеет наибольшее значение при равенстве сомножителей: так как $p+q=1$, то при $p=q=1/2$ произведение pq имеет наибольшее значение, равное $1/4$.

Пример 4.2. Из 1000 изделий, отправляемых в сборочный цех, было подвергнуто обследованию 200, отобранных случайным образом. Среди них оказалось 25 бракованных. Приняв долю бракованных изделий среди отобранных за вероятность изготовления бракованного изделия, оценить вероятность того, что во всей партии бракованных изделий окажется не более 15% и не менее 10%.

Решение. Определим вероятность изготовления бракованного изделия: $p=25/200=0{,}125$.

Наибольшее отклонение частости появлений бракованных изделий от вероятности p по абсолютной величине равно $|m/n-p|=0{,}025$; число испытаний $n=1000$. По формуле (4.15) находим искомую вероятность:

$$P\{|m/n-p|<0{,}025\}>1-(0{,}125\cdot 0{,}875)/(1000\cdot 0{,}025^2),$$

или

$$P\{|m/n-p|<0{,}025\}>0{,}825.$$

Обобщением теоремы Бернулли на случай, когда испытания происходят при неодинаковых условиях, что вызывает изменение вероятности появлений события A в каждом испытании, является теорема Пуассона.

Теорема Пуассона. *При достаточно большом числе независимых испытаний n можно с вероятностью, близкой к единице, утверждать, что разность между частостью появлений события A и средней арифметической вероятностей этого события в каждом испытании по абсолютной величине окажется меньше сколь угодно малого числа τ, если вероятность события A меняется от испытания к испытанию.* Утверждение теоремы можно записать в виде следующего неравенства:

$$P\{|m/n-\bar{p}|\leqslant\tau\}>1-\eta. \qquad (4.17)$$

где τ и η — любые сколь угодно малые положительные числа.

§ 4.5. Теорема Ляпунова

Рассмотренные теоремы закона больших чисел касаются вопросов приближения некоторых случайных величин к определенным предельным значениям независимо от их распределения. В теории вероятностей существует другая группа теорем, относящихся к предельным законам распределения суммы случайных величин. Эта группа теорем носит общее название *центральной предельной теоремы*. Различные формы центральной предельной теоремы отличаются между собой условиями, накладываемыми на сумму составляющих случайных величин.

Впервые центральная предельная теорема была сформулирована и доказана великим русским математиком А. М. Ляпуновым.

Теорема Ляпунова. *Распределение суммы независимых случайных величин X_i ($i=1, 2, ..., n$) приближается к нормальному распределению при неограниченном увеличении n, если выполняются следующие условия:*

1) *все величины имеют конечные математические ожидания и дисперсии;*

2) *ни одна из величин по своему значению резко не отличается от всех остальных.*

При решении многих практических задач используют следующую формулировку теоремы Ляпунова для средней арифметической наблюдавшихся значений случайной величины $\bar{x}$, которая также является случайной величиной (при этом соблюдаются условия, перечисленные выше): *если случайная величина X имеет конечные математическое ожидание $M(x)$ и дисперсию σ_x^2, то распределение средней арифметической $\bar{x}=(x_1+x_2+...+x_n)/n$, вычисленной по наблюдавшимся значениям случайной величины в n независимых испытаниях, при $n\to\infty$ приближается к нормальному с математическим ожиданием $M(x)$ и дисперсией σ_x^2, т. е.*

$$P\{\bar{x}<x\}\to\frac{1}{\sqrt{2\pi}\,\sigma_{\bar{x}}}\int_{-\infty}^{x} e^{-\{x-M(x)\}^2/\left(2\sigma_{\bar{x}}^2\right)}\,dx. \quad (4.18)$$

Поэтому вероятность того что $\bar{x}$ заключено в интервале (a, b), можно вычислить по формуле

$$P(a<\bar{x}<b)\approx\frac{1}{\sigma_{\bar{x}}\sqrt{2\pi}}\int_{a}^{b} e^{-\{x-M(x)\}^2\left(2\sigma_{\bar{x}}^2\right)}\,dx. \quad (4.19)$$

Используя функцию Лапласа, формулу (4.19) запишем в следующем удобном для расчетов виде:

$$P(a<\bar{x}<b)\approx{}^1/_2\,\Phi(t_2)-{}^1/_2\,\Phi(t_1),$$

где

$$t_1=(a-M(x))/\sigma_{\bar{x}};\quad t_2=(b-M(x))/\sigma_{\bar{x}}.$$

Следует отметить, что центральная предельная теорема справедлива не только для непрерывных, но и для дискретных случайных величин. Практическое значение

теоремы Лапласа огромно. Опыт показывает, что распределение суммы независимых случайных величин, сравнимых по своему рассеиванию, достаточно быстро приближается к нормальному. Уже при числе слагаемых порядка десяти распределение суммы может быть замено нормальным.

§ 4.6. Теорема Муавра—Лапласа

Частным случаем центральной предельной теоремы является теорема Муавра — Лапласа, которая известна также под названием интегральной предельной теоремы Муавра — Лапласа. В ней рассматривается случай, когда случайные величины X_i $(i=1, 2, ..., n)$ дискретны, одинаково распределены и принимают только два возможных значения: 0 и 1.

Интегральная теорема Муавра — Лапласа. *Если производится n независимых испытаний, в каждом из которых событие A появляется с вероятностью p, то для любого интервала (a; b) справедливо следующее соотношение:*

$$P(a \leqslant m \leqslant b) = P\left(\frac{a-np}{\sqrt{npq}} \leqslant \frac{m-np}{\sqrt{npq}} \leqslant \frac{b-np}{\sqrt{npq}}\right) \approx$$

$$\approx \frac{1}{2}\Phi\left(\frac{b-np}{\sqrt{npq}}\right) - \frac{1}{2}\Phi\left(\frac{a-nq}{\sqrt{npq}}\right), \qquad (4.20)$$

где m — число появлений события A в n испытаниях; $q=1-p$.

Вводя обозначения нормированных отклонений $t_1 = (a-np)/\sqrt{npq}$; $t_2=(b-np)/\sqrt{npq}$, получаем удобную для практических расчетов формулу

$$P(a \leqslant m \leqslant b) \approx {}^1/_2\Phi(t_2) - {}^1/_2\Phi(t_1). \qquad (4.21)$$

Из этой формулы с помощью преобразований можно получить следующие соотношения:

$$P\left(np + t_1\sqrt{npq} \leqslant m \leqslant np + t_2\sqrt{npq}\right) \approx$$

$$\approx {}^1/_2\Phi(t_2) - {}^1/_2\Phi(t_1), \qquad (4.22)$$

$$P\left(t_1\sqrt{npq} \leqslant m - np \leqslant t_2\sqrt{npq}\right) \approx$$

$$\approx {}^1/_2\Phi(t_2) - {}^1/_2\Phi(t_1), \qquad (4.23)$$

$$P\left(t_1\sqrt{(pq)/n} \leqslant m/n - p \leqslant t_2\sqrt{(pq/n)}\right) \approx$$

$$\approx {}^1/_2\Phi(t_2) - {}^1/_2\Phi(t_1). \qquad (4.24)$$

При $t_1 = -t_2$ формулы (4.23) и (4.24) принимают такой вид:

$$P\left(|m - np| \leqslant t\sqrt{npq}\right) \approx \Phi(t), \qquad (4.25)$$

$$P\left(|m/n - p| \leqslant t\sqrt{(pd)/n}\right) \approx \Phi(t). \qquad (4.26)$$

В отличие от неравенств (4.15) и (4.16), определяющих нижнюю границу вероятности отклонения частости появлений события от его вероятности в отдельном испытании или частоты появлений события от его математического ожидания, формулы (4.25) и (4.26) дают более точную оценку этой вероятности.

Пример 4.3. Вероятность появлений газовых раковин при отливке блока цилиндра автомобильного двигателя равна 0,2. Изготовлено 400 блоков цилиндров. Найти величину наибольшего отклонения частости отлитых блоков цилиндров с наличием газовых раковин от вероятности 0,2, которую можно гарантировать с вероятностью 0,9545.

Решение. По условию, $n=400$; $p=0{,}2$; $q=0{,}8$; $\Phi(t)=0{,}9545$. По табл. 2 приложений находим $t=2{,}0$. Используя формулу (4.26), получаем

$$P\left\{|m/n - 0{,}2| < 2\sqrt{(0{,}2\cdot 0{,}8)/400}\right\} \approx 0{,}9545,$$

или

$$P\{|m/n - 0{,}2|) < 0{,}04 \approx 0{,}9545.$$

Следовательно, величина наибольшего отклонения частости от постоянной вероятности $p=0{,}2$, которую можно гарантировать с вероятностью 0,9545, равна 0,04.

Пример 4.4. Известно, что при контроле бракуется 10% шестерен. Для контроля отобрано 500 деталей. Найти вероятность того, что число годных шестерен окажется в пределах от 460 до 475.

Решение. По условию, $n=500$; $p=0{,}9$, $q=0{,}1$; $a=460$; $b=475$. По формуле (4.21) находим искомую вероятность:

$$t_1 = (460 - 450)/\sqrt{500\cdot 0{,}9\cdot 0{,}1} = 1{,}49;$$

$$t_2 = (475 - 450)/\sqrt{500\cdot 0{,}9\cdot 0{,}1} = 3{,}73;$$

$$P(460 < m < 475) \approx {}^1/_2\,\Phi(3{,}73) - {}^1/_2\,\Phi(1{,}49) = 0{,}0679.$$

Пример 4.5. При установившемся технологическом процессе вероятность изготовления бракованных гильз равна 0,17. Для обоснования введения автоматического контроля необходимо определить количество гильз n, которое нужно направлять на контроль, чтобы с вероятностью 0,95 быть уверенным в стабильности технологического процесса, если допускается 20% брака.

Решение. По условию известно, что $m/n=0{,}2$; $p=0{,}17$; $q=0{,}83$; $\Phi(t)=0{,}95$. По табл. 2 приложений находим $t=1{,}96$. По формуле (4.26) определяем

$$P\left\{|0{,}2 - 0{,}17| < 1{,}96\sqrt{(0{,}17\cdot 0{,}83)/n}\right\} \approx 0{,}95.$$

Отсюда $n > 602$.

Глава 5

Статистическое оценивание параметров распределения

§ 5.1. Понятие об оценке параметров

При изучении качественного или количественного признака, характеризующего совокупность однородных объектов, не всегда имеется возможность обследовать каждый объект изучаемой совокупности. Приведем такой пример. Электрическую лампочку условимся считать стандартной, если продолжительность ее горения не менее 1200 ч, в противном случае она считается нестандартной. За качеством продукции обязан следить завод-изготовитель. Исследовать каждую лампочку на продолжительность горения практически невозможно, да это и противоречит здравому смыслу. Как же получить представление о качестве изготовляемой продукции? Пусть заводу необходимо поставить потребителю партию готовых изделий. Вместо данных о качестве всех электрических лампочек партии достаточно получить точные сведения о качестве небольшой их части, отобранных случайно. По продолжительности горения отобранных лампочек можно судить о качестве всех лампочек партии. Практика подтверждает, что сделанные выводы бывают достаточно надежными.

Совокупность всех возможных, иногда говорят, — всех мыслимых, значений исследуемой случайной величины называют *генеральной совокупностью.*

Множество значений случайной величины, полученное в результате наблюдений над нею, называют *случайной выборкой* или просто *выборкой.*

Число объектов в генеральной совокупности и в выборке называют их *объемами.* Генеральная совокупность может иметь как конечный, так и бесконечный объем.

В самом общем смысле содержание этой главы можно сформулировать как совокупность методов, позволяющих делать научно обоснованные выводы о числовых параметрах распределения генеральной совокупности по случайной выборке из нее. Если, например, нас интересует математическое ожидание генеральной совокупности, то задача статистической оценки параметров заключается в том, чтобы найти такую выборочную характеристику, которая позволила бы получить по возможности более точное и надежное представление об интересующем нас параметре (в данном случае о математическом ожидании). Состав выборки случаен, поэтому выводы о параметрах генеральной совокупности, сделанные по выборочным данным, могут быть ложными. С возрастанием числа элементов выборки вероятность правильного вывода увеличивается. Поэтому всякому решению, принимаемому при статистической оценке параметров, стараются поставить в соответствие вероятность, характеризующую степень достоверности принимаемого решения.

Сформулируем задачу оценки параметров в общем виде. Пусть X — случайная величина, подчиненная закону распределения $F(X, \Theta)$, где Θ — параметр распределения, числовое значение которого неизвестно. Исследовать все элементы генеральной совокупности для вычисления параметра Θ не представляется возможным, поэтому об этом параметре пытаются судить по выборкам из генеральной совокупности.

Всякую однозначно определенную функцию результатов наблюдений, с помощью которой судят о значении параметра Θ, называют *оценкой* (или *статистикой*) параметра Θ. Рассмотрим некоторое множество выборок объемом n каждая. Выборочную оценку параметра Θ, вычисленную по i-й выборке, обозначим Θ_{n_i}. Так как состав выборки случаен, то можно сказать, что Θ_{n_i} примет неизвестное заранее числовое значение, т. е. является случайной величиной.

Известно, что случайная величина определяется законом распределения и числовыми характеристиками, следовательно, и выборочную оценку можно также описывать законом распределения и числовыми характеристиками.

Для того чтобы отразить случайный характер выборки объема n из генеральной совокупности, обозначим ее

через $(X_1, X_2, ..., X_n)$, а выборочную оценку параметра Θ — через $\tilde{\Theta}_n$ (иногда выборочную оценку обозначают $\hat{\Theta}_n$). Следовательно, можно записать $\tilde{\Theta}_n = f(X_1, X_2, ..., X_n)$. Выбор оценки, позволяющей получить хорошее приближение оцениваемого параметра, — основная задача теории оценивания.

§ 5.2. Основные свойства оценок

Основной задачей теории оценивания параметров, как было изложено выше, является получение научно обоснованных выводов о параметрах генеральной совокупности по выборке. В этом параграфе сформулированы свойства, которым должна отвечать выборочная оценка. Основными свойствами оценок являются свойства несмещенности, эффективности и состоятельности.

Оценку $\tilde{\Theta}_n$ параметра Θ называют *несмещенной*, если ее математическое ожидание равно оцениваемому параметру Θ, т. е. $M(\tilde{\Theta}_n) = \Theta$. Если это равенство не выполняется, то оценка $\tilde{\Theta}_n$ может либо завышать значение Θ (т. е. $M(\tilde{\Theta}_n) > \Theta$), либо занижать его (т. е. $M(\tilde{\Theta}_n) < \Theta$). В обоих случаях это приводит к систематическим (одного знака) ошибкам в оценке параметра Θ. Требование несмещенности гарантирует отсутствие систематических ошибок при оценке параметров. Так как $\tilde{\Theta}_n$ — случайная величина, значение которой изменяется от выборки к выборке, то меру ее рассеивания около математического ожидания Θ будем характеризовать дисперсией $D(\tilde{\Theta}_n)$. Пусть $\tilde{\Theta}_n$ и $\tilde{T}_n$ — две несмещенные оценки параметра Θ, т. е. $M(\tilde{\Theta}_n) = \Theta$ и $M(\tilde{T}_n) = \Theta$, причем дисперсии $\tilde{\Theta}_n$ и T_n обозначим соответственно $D(\tilde{\Theta}_n)$ и $D(\tilde{T}_n)$. Из двух оценок $\tilde{\Theta}_n$ и $\tilde{T}_n$ следует отдать предпочтение той, которая обладает меньшим рассеиванием около оцениваемого параметра, так как в этом случае значение оценки, вычисленное по конкретной выборке, меньше всего отличается от истинного значения оцениваемого параметра. Если $D(\tilde{\Theta}_n) < D(\tilde{T}_n)$, то в качестве оценки Θ принимают $\tilde{\Theta}_n$.

Несмещенную оценку $\tilde{\Theta}_n$, которая имеет наименьшую дисперсию среди всех возможных несмещенных оценок параметра Θ, вычисленных по выборкам одного и того же объема, называют *эффективной оценкой.*

Оценку $\tilde{\Theta}_n$ параметра Θ называют *состоятельной,* если она подчиняется закону больших чисел, т. е. при достаточно большом числе независимых наблюдений n с вероятностью, близкой к единице, можно утверждать, что разность между $\tilde{\Theta}_n$ и Θ по абсолютной величине окажется меньше сколь угодно малого положительного числа τ, или

$$P\{|\tilde{\Theta}_n - \Theta| < \tau\} > 1 - \eta,$$

где η — положительное число, близкое к нулю. Следовательно, в случае использования состоятельных оценок оправдывается увеличение числа единиц выборки, так как при этом все менее вероятной становится возможность значительной ошибки в оценке неизвестного параметра.

На практике при оценке параметров не всегда удается удовлетворить одновременно требованиям несмещенности, эффективности и состоятельности оценки. Так, например, может оказаться, что для простоты расчетов целесообразно использовать незначительно смещенную оценку. Однако выбору оценки всегда должно предшествовать ее критическое рассмотрение со всех точек зрения.

§ 5.3. Оценка математического ожидания и дисперсии по выборке

Наиболее важными числовыми характеристиками случайной величины являются математическое ожидание и дисперсия.

Рассмотрим вопрос о том, какие выборочные характеристики лучше всего в смысле несмещенности, эффективности и состоятельности оценивают математическое ожидание и дисперсию.

Теорема 5.1. *Средняя арифметическая $\bar{X}$, вычисленная по n независимым наблюдениям над случайной величиной X, которая имеет математическое ожидание μ, является несмещенной оценкой этого параметра.*

Доказательство. Пусть X_1, X_2, ..., X_i, ..., X_n — n независимых наблюдений над случайной величиной X. По условию, случайная величина X имеет математическое ожидание, равное μ, а это значит, что $M(X_i)=\mu$; $i=\overline{1, n}$. По определению, средняя арифметическая

$$\overline{X}=\sum_{i=1}^{n} X_i/n.$$

Рассмотрим теперь математическое ожидание средней арифметической. Используя свойства математического ожидания, имеем

$$M(\overline{X})=\sum_{i=1}^{n} M(X_i)/n=\sum_{i=1}^{n} \mu/n=n\mu/n=\mu,$$

следовательно, $M(\overline{X})=\mu$.

Теорема 5.2. *Средняя арифметическая $\overline{X}$, вычисленная по n независимым наблюдениям над случайной величиной X, которая имеет математическое ожидание μ и дисперсию σ^2, является состоятельной оценкой этого параметра.*

Доказательство. Пусть X_i ($i=1, 2, ..., n$) — независимые наблюдения над случайной величиной X. Условия предыдущей теоремы сохраняются и в этом случае, поэтому $M(\overline{X})=\mu$. Запишем неравенство Чебышева для средней арифметической $\overline{X}$:

$$P\{|\overline{X}-\mu|<\tau\}>1-\sigma^2(\overline{X})/\tau^2,$$

$$P\{|\overline{X}-\mu|\geqslant\tau\}\leqslant\frac{D(\overline{X})}{\tau^2}.$$

Известно, что $D(\overline{X})=\Sigma D(X_i)/n^2$. По условию теоремы, X — случайная величина, дисперсия которой σ^2, следовательно, $D(X_i)=\sigma^2$. Таким образом,

$$D(\overline{X})=1/n^2(\underbrace{\sigma^2+...+\sigma^2+...+\sigma^2}_{n})=(1/n^2)(n\sigma^2)=\sigma^2/n.$$

Итак, доказано, что дисперсия средней арифметической в n раз меньше дисперсии случайной величины X. Тогда

$$\sigma^2(\overline{X})/\tau^2=\sigma^2/(n\tau^2)=\eta$$

при достаточном большом n является числом, близким к нулю. Поэтому для сколь угодно малого положительного числа τ выполняется неравенство

$$P\{|\overline{X}-\mu|<\tau\}>1-\eta,$$

определяющее свойство состоятельности выборочных оценок.

Получение эффективных оценок — сложное дело, поэтому отсылаем заинтересованного читателя к специальным курсам математической статистики. Приведем без доказательства важный для практики результат. Если случайная величина X распределена по нормальному закону с параметрами (μ, σ^2), то несмещенная оценка $\overline{X}$ математического ожидания μ имеет минимальную дисперсию, равную σ^2/n, поэтому средняя арифметическая $\overline{X}$ в этом случае является эффективной оценкой математического ожидания μ.

Докажем некоторые результаты, относящиеся к оценке дисперсии.

Теорема 5.3. *Если случайная выборка состоит из n независимых наблюдений над случайной величиной X с математическим ожиданием μ и дисперсией σ^2, то выборочная дисперсия*

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

не является несмещенной оценкой генеральной дисперсии.

Доказательство. Пусть X_i $(i=1, 2, \ldots, n)$ — независимые наблюдения над случайной величиной X. По условию,

$$M(X_1)=M(X_2)=\ldots=M(X_i)=\ldots=M(X_n)=\mu,$$
$$D(X_1)=D(X_2)=\ldots=D(X_i)=\ldots=D(X_n)=\sigma^2.$$

Преобразуем формулу выборочной дисперсии:

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu+\mu-\overline{X})^2=$$

$$=\frac{1}{n}\sum_{i=1}^{n}[(X_i-\mu)-(\overline{X}-\mu)]^2=$$

$$= {}^1/_n \sum_{i=1}^{n}(X_i-\mu)^2 - {}^2/_n (\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + {}^n/_n (\overline{X}-\mu)^2.$$

Упростим выражение $^2/_n(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)$. Принимая во внимание, что $\overline{X}=(\sum_{i=1}^{n} X_i)/n$, откуда $n\overline{X}=\sum_{i=1}^{n} X_i$, можно записать

$$\frac{2(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)}{n} = \frac{2(\overline{X}-\mu)\left(\sum_{i=1}^{n} X_i - n\mu\right)}{n} =$$

$$= \frac{2(\overline{X}-\mu)(n\overline{X}-n\mu)}{n} = 2(\overline{X}-\mu)^2.$$

Тогда

$$s^2 = \left(\sum_{i=1}^{n}(X_i-\mu)^2\right)\Big/ n - 2(\overline{X}-\mu)^2 + (\overline{X}-\mu)^2 =$$

$$= \left(\sum_{i=1}^{n}(X_i-\mu)^2\right)\Big/ n - (\overline{X}-\mu)^2.$$

Теперь рассмотрим математическое ожидание от выборочной дисперсии:

$$M(s^2) = M\left[\left(\sum_{i=1}^{n}(X_i-\mu)^2\right)\Big/ n\right] - M(\overline{X}-\mu)^2.$$

Используя определение и свойства дисперсии можно доказать, что

$$M\left[\left(\sum_{i=1}^{n}(X_i-\mu)^2\right)\Big/ n\right] = \sigma^2 \text{ и } M(\overline{X}-\mu)^2 =$$

$$= D(\overline{X}) = \sigma^2/n,$$

следовательно, $M(s^2)=\sigma^2-\sigma^2/n$. Итак,

$$M(s^2) = \sigma^2 - \sigma^2/n = (n-1)\sigma^2/n,$$

т. е. s^2 является смещенной оценкой дисперсии генеральной совокупности. □

Приведем несмещенную оценку дисперсии генеральной совокупности

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Обычно эту оценку называют *исправленной выборочной дисперсией.* Действительно,

$$\hat{s}^2 = \frac{n}{n-1}s^2,$$

тогда

$$M\left(\hat{s}^2\right) = M\left(\frac{n}{n-1}s^2\right) = \frac{n}{n-1}M(s^2) =$$

$$= \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2.$$

Дробь $n/(n-1)$ называют поправкой Бесселя. При малых значениях n поправка Бесселя довольно значительно отличается от единицы, с увеличением же n она быстро стремится к единице. При $n > 50$ практически нет разницы между оценками s^2 и $\hat{s}^2$. Можно показать, что оценки s^2 и $\hat{s}^2$ являются состоятельными оценками σ^2.

Несмещенной, состоятельной и эффективной оценкой σ^2 является оценка

$$s_*^2 = \left(\sum_{i=1}^{n}(X_i - \mu)^2\right)\Big/ n,$$

для вычисления которой необходимо знать математическое ожидание. Заметим, что оценка s_*^2 эффективна лишь при условии нормальности распределения случайной величины X в генеральной совокупности. Оценки s^2 и $\hat{s}^2$ не являются эффективными. В том случае, когда значение математического ожидания неизвестно, для оценки дисперсии σ^2 используют состоятельную и несмещенную оценку $\hat{s}^2$.

§ 5.4. Метод моментов

В § 5.2 были сформулированы основные свойства оценок, но ничего не говорилось о способах их получения. Одним из первых методов оценивания параметров был метод моментов, разработанный К. Пирсоном.

Пусть известный закон распределения случайной величины X определяется несколькими параметрами Θ_1, Θ_2, ..., Θ_q, числовое значение которых неизвестно. Для нахождения выборочных оценок производят n независимых наблюдений x_i $(i=\overline{1, n})$ над случайной величиной X. Следуя методу моментов, необходимо q первых моментов случайной величины X приравнять q выборочным моментам, полученным из экспериментальных данных. Формулы для расчета начальных и центральных моментов по выборке приведены в § 1.12.

Теоретическим обоснованием метода моментов служит закон больших чисел, согласно которому для рассматриваемого случая при большом объеме выборки выборочные моменты близки к моментам в генеральной совокупности.

Метод моментов дает возможность получать состоятельные оценки. Таким образом, надежность выводов, сделанных при его использовании, зависит от количества наблюдений.

В качестве примера применения метода моментов рассмотрим случайную величину X, имеющую нормальный закон распределения, который определяется двумя параметрами: математическим ожиданием $M(X)$ и дисперсией $D(X)$. Известно, что $M(X)$ — начальный момент первого порядка (ν_1), а $D(X)$ — центральный момент второго порядка, (μ_2).

Согласно § 1.9, ν_1 можно оценить эмпирическим начальным моментом первого порядка, а μ_2 — эмпирическим центральным моментом второго порядка. Таким образом, по методу моментов неизвестное математическое ожидание оценивается средним арифметическим, а дисперсия — выборочной дисперсией s^2.

Оцениваемые параметры также могут быть выражены в виде определенных функций теоретических моментов. Так, например, эксцесс $E=\mu_4/\mu_2^2-3$. Заменяя теоретические моменты моментами выборочного распределения, можно и в этом случае получить выборочную оценку неизвестного параметра. В рассматриваемом примере эксцесс по выборке оценивается так: $\tilde{E}=\tilde{\mu}_4/s^4-3$. Использование метода моментов на практике приводит к сравнительно простым вычислениям.

Исследуя свойства оценок, получаемых с помощью метода моментов, английский математик Р. А. Фишер

предложил более надежный метод оценивания параметров генеральной совокупности — метод максимального правдоподобия. Хотя метод максимального правдоподобия часто приводит к более сложным вычислениям, чем метод моментов, все же оценки, получаемые с его помощью, как правило, оказываются более надежными и особенно предпочтительными в случае малого числа наблюдений.

§ 5.5. Метод максимального правдоподобия

Основным способом получения оценок параметров генеральной совокупности по данным выборки является метод максимального правдоподобия. Основная идея метода заключается в следующем.

Пусть $X_i (i = \overline{1, n})$ — результаты n независимых наблюдений над случайной величиной X, которая может быть как дискретной, так и непрерывной; $f(X, \Theta)$ — вероятность значения (если случайная величина дискретна) и плотность вероятности (если случайная величина непрерывна). Функция $f(X, \Theta)$ зависит от неизвестного параметра Θ, который требуется оценить по выборке.

Если $X_1, X_2, \ldots, X_n$ — независимые случайные величины, то функцией правдоподобия называется выражение

$$L = f(X_1, \Theta) f(X_2, \Theta) \ldots f(X_n, \Theta). \qquad (5.1)$$

В качестве оценки неизвестного параметра Θ берется такое значение $\tilde{\Theta}$, при подстановке которого в выражение (5.1) вместо параметра Θ получаем максимальное значение функции L. Оценку $\tilde{\Theta}$ обычно называют *оценкой максимального правдоподобия*. Заметим, что оценка $\tilde{\Theta}$ зависит от количества и числовых значений случайных величин X_i, следовательно, сама является случайной величиной.

При максимизации функции L подразумевается, что значения $X_1, X_2, \ldots, X_n$ фиксированы, а переменной величиной является параметр $\tilde{\Theta}$ (иными словами, максимум L отыскивается в предположении, что X_i заменены их числовыми значениями). Если L дифференцируема относительно параметра Θ, то для отыскания максимума функции необходимо воспользоваться известными правилами определения экстремальных значений функции,

т. е. для рассматриваемого случая решить уравнение

$$\frac{\partial L}{\partial \Theta} = 0 \qquad (5.2)$$

и в качестве оценки Θ выбрать решение, которое обращает функцию L в максимум. Если функция L недифференцируема относительно параметра Θ, то для вычисления ее максимума приходится использовать другие методы, что иногда связано с большими вычислительными трудностями. Иногда вместо уравнения (5.2) удобнее рассматривать уравнение

$$\frac{1}{L}\frac{\partial L}{\partial \Theta} = \frac{\partial \ln L}{\partial \Theta} = 0. \qquad (5.3)$$

Если закон распределения случайной величины зависит от нескольких параметров $\Theta_1, \Theta_2, ..., \Theta_k$, то ход рассуждений аналогичен предыдущему, т. е. необходимо найти такие величины $\tilde{\Theta}_1, \tilde{\Theta}_2, .., \tilde{\Theta}_k$, которые максимизировали бы функцию правдоподобия. Если эта функция дифференцируема относительно каждого из оцениваемых параметров, то отыскание оценок наибольшего правдоподобия сводится к отысканию максимума функции L относительно параметров $\Theta_1, \Theta_2, ..., \Theta_k$ по правилу определения экстремума функции нескольких переменных.

Рассмотрим пример, иллюстрирующий общие идеи метода максимального праводоподобия. Пусть x_1, $x_2, ..., x_n$—n независимых наблюдений над случайной величиной X, подчиняющейся нормальному закону распределения.

Как известно, нормальный закон распределения определяется двумя параметрами: математическим ожиданием μ и дисперсией σ^2. Функция $f(X, \Theta)$* в нашем случае имеет вид

$$f(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

Задача заключается в том, чтобы по данным выборки оценить неизвестные параметры распределения μ и σ^2. Функцию наибольшего правдоподобия можно записать в следующем виде:

---

* Под Θ в данном случае следует подразумевать вектор $\Theta = (\mu, \sigma^2)$.

$$L = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_1-\mu)^2/(2\sigma^2)} \frac{1}{\sigma\sqrt{2\pi}} \times$$

$$\times e^{-(x_2-\mu)^2/(2\sigma^2)} \ldots \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_n-\mu)^2/(2\sigma^2)} =$$

$$= \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Функция L дифференцируема относительно параметров μ и σ^2. Для упрощения выкладок имеет смысл искать максимум не самой функции L, а $\ln L$, поскольку и L и $\ln L$ имеют максимум в одной и той же точке:

$$L = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2},$$

$$\ln L = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \ln e.$$

Так как $\ln e = 1$, то

$$\ln L = n \ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2. \qquad (5.4)$$

Для нахождения оценок параметров μ и σ^2 необходимо решить совместно следующую систему уравнений:

$$\begin{cases} \dfrac{\partial \ln L}{\partial \mu} = 0, \\ \dfrac{\partial \ln L}{\partial \sigma^2} = 0. \end{cases} \qquad (5.5)$$

Найдем первую производную от $\ln L$ по параметру μ. Заметим, что $\partial\left(n \ln \dfrac{1}{\sigma\sqrt{2\pi}}\right)\Big/ \partial\mu = 0$, так как при дифференцировании по μ выражение $n\ln(1/(\sigma\sqrt{2\pi}))$ является константой. Следовательно,

$$\frac{\partial \ln L}{\partial \mu} = \frac{\partial\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]}{\partial \mu} =$$

$$= -\frac{1}{2\sigma^2}\,\frac{\partial\left([x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_i-\mu)^2+\ldots+(x_n-\mu)\right]^2}{\partial \mu} =$$

$$= -\frac{1}{2\sigma^2}\left[-2(x_1-\mu)-2(x_2-\mu)-\ldots-2(x_i-\mu)-\ldots\right.$$

$$\left.\ldots-2(x_n-\mu)\right] = \frac{1}{2\sigma^2}\cdot 2\sum_{i=1}^{n}(x_i-\mu) = \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2}.$$

Итак, $\frac{\partial \ln L}{\partial \mu} = \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2}$.

Для определения оценки μ частную производную от $\ln L$ по параметру μ приравняем нулю:

$$\left(\sum_{i=1}^{n}(x_i-\mu)\right)\Big/\sigma^2 = 0,$$

$$\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\mu = 0,\ \sum_{i=1}^{n}x_i - n\mu = 0,$$

следовательно, $n\mu = \sum_{i=1}^{n}x_i$, откуда $\mu = (\sum_{i=1}^{n}x_i)/n$, но $(\sum_{i=1}^{n}x_i)/n$, по определению есть $\bar{x}$. Таким образом, математическое ожидание следует оценивать с помощью средней арифметической.

Исследуем вторую частную производную:

$$\frac{\partial^2 \ln L}{\partial \mu^2} = \frac{\partial\left[\sum_{i=1}^{n}(x_i-\mu)/\sigma^2\right]}{\partial \mu} = \frac{1}{\sigma^2}\times$$

$$\times\,\frac{\partial\left[(x_1-\mu)+(x_2-\mu)+\ldots+(x_n-\mu)\right]}{\partial \mu} =$$

$$= \frac{1}{\sigma^2}\left(\frac{\partial x_1}{\partial \mu}+\frac{\partial x_2}{\partial \mu}+\ldots+\frac{\partial x_i}{\partial \mu}+\ldots+\frac{\partial x_n}{\partial \mu}-n\frac{\partial \mu}{\partial \mu}\right) =$$

$$= \frac{1}{\sigma^2}\underbrace{(0+0+\ldots+0+\ldots 0-n)}_{n} = -\frac{n}{\sigma^2} < 0.$$

Далее перейдем к преобразованиям, связанным со вторым уравнением системы (5.5). Найдем первую производную $\ln L$ по параметру σ^2:

$$\frac{\partial \ln L}{\partial (\sigma^2)} = \frac{\partial \left[n \ln 1/\sqrt{2\pi} \cdot 1/\sqrt{\sigma^2} - 1/2\sigma^2 \sum_{i=1}^{n} (x_i - \mu)^2 \right]}{\partial (\sigma^2)} =$$

$$= \frac{n\partial \left[\ln 1 - \ln \sqrt{2\pi} - 1/2 \ln (\sigma^2) \right]}{\partial (\sigma^2)} - \frac{\partial \left[1/2\sigma^2 \sum_{i=1}^{n} (x_i - \mu)^2 \right]}{\partial (\sigma^2)} =$$

$$= -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^4}.$$

Для определения оценки σ^2 производную от $\ln L$ по параметру σ^2 приравняем нулю:

$$-\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^4} = 0 \quad (\sigma^2 \neq 0),$$

или

$$-n\sigma^2 + \sum_{i=1}^{n} (x_i - \mu)^2 = 0,$$

откуда $\sigma^2 = \sum_{i=1}^{n} (x_i - \mu^2)/n$. Учитывая, что $\bar{X} = \mu$, можно записать

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/n.$$

Вторая частная производная также отрицательна.

Анализируя результаты, полученные при решении системы (5.5), можно сказать, что точка с координатами $(\mu = \bar{x};\ \sigma^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2)/n$ максимизирует значение функции $\ln L$, следовательно и саму функцию L, так как в экстремальной точке определитель

$$Q = \begin{vmatrix} \frac{\partial^2 \ln L}{\partial \mu^2} & \frac{\partial^2 \ln L}{\partial \mu \partial (\sigma^2)} \\ \frac{\partial^2 \ln L}{\partial \mu \partial (\sigma^2)^2} & \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} \end{vmatrix} > 0.$$

Действительно,

$$\begin{vmatrix} -n/\sigma^2 & -\sum_{i=1}^{n}(x_i-\mu)\cdot 1/\sigma_4 \\ -\sum_{i=1}^{n}(x_i-\mu)\cdot 1/\sigma^4 & -n/2\sigma^4 \end{vmatrix} = n^2/2\,(\sigma^2)^3 - \left[\sum_{i=1}^{n}(x_i-\mu)\right]^2 1/(\lambda^2)^4.$$

Выражение $\sum_{i=1}^{n}(x_i-\bar{x}) = \sum_{i=1}^{n}x_i - n\bar{x}=0$, поэтому, так как $n^2>0$, $\sigma^2>0$, значение определителя больше нуля: $Q=n^2/[2(\sigma^2)^6]>0$. Итак, если случайная величина подчинена нормальному закону распределения, то по данным выборки математическое ожидание следует оценивать с помощью средней арифметической, а дисперсию генеральной совокупности — дисперсией выборки s^2. Отвечают ли полученные оценки основным свойствам оценок, т. е. несмещенности, эффективности и состоятельности? Как следует из § 5.3, средняя арифметическая, вычисленная по выборке, которая взята из совокупности с нормальным законом распределения, является несмещенной, эффективной и состоятельной оценкой математического ожидания.

Метод максимального правдоподобия для оценки математического ожидания также предполагает использование средней арифметической. Однако, применяя метод максимального правдоподобия, не всегда можно получить оценки, удовлетворяющие одновременно всем трем критериям.

Так, например, дисперсию генеральной совокупности в вышеразобранном примере по методу максимального правдоподобия рекомендуется оценивать выборочной дисперсией s^2. Как следует из § 5.3, выборочная дисперсия s^2 удовлетворяет лишь условию состоятельности. Поэтому к каждой оценке, получаемой с помощью метода максимального правдоподобия, следует подходить критически, т. е. всякий раз анализировать, удовлетворяет ли она основным свойствам оценок.

Заметим, что иногда не существует оценок, одновременно удовлетворяющих свойствам несмещенности, эффективности и состоятельности. Если математическое

ожидание генеральной совокупности неизвестно, не существует эффективной оценки дисперсии генеральной совокупности. В подобных случаях надо стремиться к тому, чтобы получить оценку, удовлетворяющую по возможности наибольшему количеству свойств, предъявляемых к оценкам.

Например, дисперсию генеральной совокупности при неизвестном математическом ожидании следует оценивать исправленной выборочной дисперсией $\hat{s}^2$, которая удовлетворяет свойствам и несмещенности и состоятельности, а не s^2, которая является лишь состоятельной оценкой генеральной дисперсии.

Метод максимального правдоподобия нашел большое применение благодаря следующим свойствам.

1°. *Закон распределения выборочной оценки* $\tilde{\Theta}_n$, *являющейся решением уравнения* (5.2), *имеет в пределе* (*при* $n\to\infty$) *нормальный закон распределения, или, как говорят* $\tilde{\Theta}_n$ *отвечает свойству асимптотической нормальности.*

2°. *Выборочная оценка* $\tilde{\Theta}_n$ *при* $n\to\infty$ *отвечает также свойству асимптотической несмещенности и асимптотической эффективности.*

Метод моментов этими свойствами не обладает и менее предпочтителен.

В заключение покажем, что s^2 — асимптотически несмещенная оценка σ^2. Свойство асимптотической несмещенности оценки $\tilde{\Theta}_n$ можно записать так: $|M(\tilde{\Theta}_n)-\Theta|\to 0$ при $n\to\infty$. По теореме 5.3, $M(s^2)=\sigma^2-\sigma^2/n$, следовательно, $|M(s^2)-\sigma^2|=\sigma^2/n\to 0$ при $n\to\infty$.

§ 5.6. Метод наименьших квадратов

В § 5.5 изложен метод получения оценок параметров генеральной совокупности по выборке — метод максимального правдоподобия, который всегда приводит к состоятельным оценкам, хотя иногда и смещенным. Известно, что этот метод использует наилучшим образом всю информацию о неизвестном параметре, содержащуюся в выборке. Однако часто его применение связано с необходимостью решения сложных систем уравнений.

Другим способом, имеющим большое практическое применение в задачах оценивания неизвестных параметров генеральной совокупности по выборке и часто при-

водящим к более простым выкладкам, является *метод наименьших квадратов*. Основоположниками этого метода являются Лежандр, Р. Адрейн, Гаусс.

Пусть, как и прежде, X — случайная величина (дискретная или непрерывная) с законом распределения $f(X, \Theta)$, (где Θ — неизвестный параметр генеральной совокупности, который следует оценить по выборке; X_1, X_2, ..., X_n — n независимых наблюдений, а $\tilde{\Theta}$ — оценка параметра Θ, зависящая от количества наблюдений и их числовых значений, т. е. $\tilde{\Theta}=\tilde{\Theta}(x_i)$.

Основная идея метода наименьших квадратов в приложении к оцениванию параметров сводится к тому, чтобы в качестве оценки неизвестного параметра принимать значение, которое минимизирует сумму квадратов отклонений между оценкой и параметром для всех наблюдений, т. е.

$$\sum_{i=1}^{n}[\Theta-\tilde{\Theta}(x_i)]^2=\min.$$

Если X — случайная величина, имеющая нормальный закон распределения $f(x, \mu, \sigma^2)$, причем σ^2 известно, а $x_1, x_2,\ldots,x_n$ — независимые наблюдения над случайной величиной X, то для оценки μ, следуя методу максимального правдоподобия, необходимо максимизировать

$$L=\frac{1}{(\sigma\sqrt{2\pi})^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2},$$

что приводит к минимизации $\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$ или $\sum_{i=1}^{n}(x_i-\mu)^2$, так как σ^2 — константа.

Выражение $\sum_{i=1}^{n}(x_i-\mu)^2=\min$ является требованием метода наименьших квадратов; используя его, можно определить выборочную оценку неизвестного параметра μ. Если исходная случайная величина имеет нормальный закон распределения, то метод максимального правдоподобия и метод наименьших квадратов дают одинаковые результаты.

Особенно часто метод наименьших квадратов применяют в задачах выравнивания или сглаживания. Пусть в результате наблюдений получен ряд точек с координатами $(x_1; y_1)$, $(x_2; y_2)$, ..., $(x_n; y_n)$ (рис. 5.1). Если заранее известно, что зависимость между переменными линейная: $Y_т = a_0 + a_1 x$ (разброс объясняется действием ненаблюдаемых факторов, ошибками измерений и т. д.), то необходимо определить числовые параметры (a_0 и a_1), которые наилучшим образом, в смысле метода наименьших квадратов, описывали бы зависимость, полученную при наблюдении. Наилучшее согласование достигается в случае, когда сумма квадратов отклонений опытных точек от точек, рассчитанных по теоретической кривой, обращается в минимум:

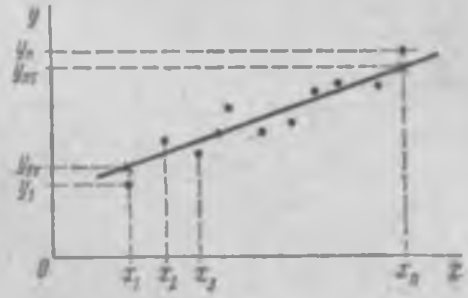

Рис. 5.1

$$S = \sum_{i=1}^{n} (Y_i - Y_{iт})^2 = \min.$$

Так как, по условию, $Y_{iт} = a_0 + a_1 x_i$, то

$$S = \sum_{i=1}^{n} (Y_i - a_0 - a_1 x_i)^2 = \min.$$

Функция S — функция двух независимых переменных a_0 и a_1. Для определения экстремума (максимума или минимума) функции нескольких переменных необходимо обращение в нуль ее частных производных первого порядка. Это условие, примененное к S, приводит к уравнениям

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = 2\left[\sum\limits_{i=1}^{n} (Y_i - a_0 - a_1 x_i)(-1)\right] = 0, \\ \dfrac{\partial S}{\partial a_1} = 2\left[\sum\limits_{i=1}^{n} (Y_i - a_0 - a_1 x_i)(-x_i)\right] = 0. \end{cases}$$

Сокращая все члены уравнений на 2 и группируя члены, содержащие a_0 и a_1, получаем

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i, \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i. \end{cases}$$

Теперь для определения a_0 и a_1 необходимо по существующим правилам решить систему двух уравнений с двумя неизвестными. В соответствующей литературе доказывается, что решение системы уравнений, если оно существует, определяет значения параметров, минимизирующих S.

Пример 5.1. Описать линейной зависимостью соотношение между урожайностью Y (в центнерах с гектара) и количеством внесенных удобрений X (в центнерах на гектар).

x_i	10	30	50	70
y_i	11	13	16	18

Решение. При $n=4$ имеем:

$$\sum_{i=1}^{4} x_i = 10 + 30 + 50 + 70 = 160;$$

$$\sum_{i=1}^{4} y_i = 11 + 13 + 16 + 18 = 58; \quad \sum_{i=1}^{4} x_i^2 = 10^2 + 30^2 + 50^2 + 70^2 = 8400;$$

$$\sum_{i=1}^{4} x_i y_i = 10 \cdot 11 + 30 \cdot 13 + 50 \cdot 16 + 70 \cdot 18 = 2560;$$

Составляем систему уравнений:

$$\begin{cases} 4a_0 + 160a_1 = 58, \\ 160a_0 + 8400a_1 = 2560. \end{cases}$$

Решением системы являются $a_0=9{,}7$, $a_1=0{,}12$. Тогда $Y_x=9{,}7+0{,}12X$.

§ 5.7. Распределение средней арифметической для выборок из нормальной совокупности. Распределение Стьюдента

Выборочная средняя, вычисленная по конкретной выборке, есть определенное число. Так как состав выборки случаен, то средняя арифметическая, вычисленная для элементов другой выборки того же объема из той же генеральной совокупности, определяется числом, как правило, отличным от первого, т. е. средняя изменяется от выборки к выборке.

Следовательно, выборочную среднюю можно рассматривать как случайную величину, что позволяет говорить о законе распределения выборочной средней. Приведем без доказательства следующую теорему.

Теорема 5.4. *Если случайная величина X подчиняется нормальному закону распределения с параметрами* (μ, σ^2), *а* $X_1, X_2, \ldots, X_n$ — *ряд независимых наблюдений над случайной величиной* X, *каждое из которых имеет те же числовые характеристики, что и* X, *т. е.* $M(X_i)=\mu$, $D(X_i)=\sigma^2$, *то выборочная средняя* $\overline{X}=\sum_{i=1}^{n} X_i/n$ *также подчиняется нормальному закону распределения.*

Как следует из § 5.3, $M(\overline{X})=\mu$, $D(\overline{X})=\sigma^2/n$. Поэтому нормированное отклонение $(\overline{X}-\mu)/(\sigma/\sqrt{n})$ подчиняется нормальному закону распределения со средним значением, равным нулю, и дисперсией, равной единице. Действительно, используя свойства математического ожидания, а также тот факт, что $\overline{X}$ и μ независимы, имеем:

$$M\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)=\frac{\sqrt{n}}{\sigma}M(\overline{X}-\mu)=\frac{\sqrt{n}}{\sigma}\left[M(\overline{X})-M(\mu)\right]=$$

$$=(\sqrt{n}/\sigma)(\mu-\mu)=0,$$

$$D\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)=\frac{n}{\sigma^2}\left[D(\overline{X})-D(\mu)\right]=\frac{n}{\sigma^2}D(\overline{X})=$$

$$=\frac{n}{\sigma^2}\frac{\sigma^2}{n}=1.$$

Пример 5.2. Автомат штампует детали. Контролируется длина детали, которая подчиняется нормальному закону распределения. Найти вероятность того, что средняя длина $\overline{x}$ деталей, отобранных

случайным образом, отклонится от математического ожидания более чем на 2 мм, если дисперсия случайной величины X равна $\sigma^2 = 9$ мм^2, а количество деталей в выборке $n = 16$.

Решение. Обозначим математическое ожидание случайной величины μ. Случайная величина $\overline{X}$ имеет нормальное распределение с математическим ожиданием μ и дисперсией $\sigma^2_{\overline{X}} = \sigma^2/n = 9/16$ мм^2, или $\sigma_{\overline{X}} = 3/4$ мм^2. Теперь найдем вероятность того, что $|\overline{X} - \mu| \geqslant 2$:

$$P(|\overline{X} - \mu| < 2) = P[-2 < (\overline{X} - \mu) < 2] =$$

$$= P(\underbrace{\mu - 2}_{a} < \overline{X} < \underbrace{\mu + 2}_{b}) = {}^1/_2 \left[\Phi\left(\frac{b-\mu}{\sigma_{\overline{X}}}\right) - \right.$$

$$\left. - \Phi\left(\frac{a-\mu}{\sigma_{\overline{X}}}\right)\right] = {}^1/_2 \left[\Phi\left(\frac{\mu+2-\mu}{3/4}\right) - \Phi\left(\frac{\mu-2-\mu}{3/4}\right)\right] =$$

$$= {}^1/_2 [\Phi(8/3) - \Phi(-8/3)] = {}^1/_2 [\Phi(8/3) + \Phi(8/3)[=$$

$$= \Phi(8/3) = 0{,}9922;$$

следовательно, $P(|\overline{X} - \mu| \geqslant 2) = 1 - 0{,}9922 = 0{,}0078$, т. е. практически можно быть уверенным, что наблюдаемая средняя длина детали отклонится от заданной, принимаемой за математическое ожидание, не более чем на 2 мм.

Итак, если случайная величина X имеет нормальный закон распределения, то нормированное отклонение $z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$ также подчиняется нормальному закону распределения. Однако дисперсия генеральной совокупности σ^2 почти всегда оказывается неизвестной, поэтому вызывает большой практический интерес изучение распределения статистики $t = \frac{\overline{X} - \mu}{\hat{s}^2}\sqrt{n}$, где $\hat{s}^2$ — несмещенная и самостоятельная оценка дисперсии σ^2, вычисленная по выборочным данным. В литературе по математической статистике доказывается, что распределение статистики t не зависит ни от математического ожидания μ случайной величины X, ни от дисперсии σ^2, а зависит лишь от объема выборки n. Закон распределения статистики t называют *t-распределением Стьюдента*. Для плотности распределения Стьюдента $s(t, n)$ составлена табл. 6 приложений. Функция $s(t, n)$ является четной, т. е. $s(-t, n) = s(t, n)$.

Рассмотрим правило пользования таблицей t-распределения. Для всякого фиксированного значения n вычисляют $k = n - 1$, которому отвечают значения $t_{k,p}$, каждое

из которых определяется выбранной вероятностью,

$$P = \int_{-t}^{t} s(t, k)\,dt.$$

Так, $n=13$ соответствуют следующие значения распределения Стьюдента: $t_{12;\,0,95}=2,18$; $t_{12;\,0,99}=3,06$; $t_{12;\,0,999}=4,32$. Заметим, что существуют более подробные таблицы изменения вероятности P и объема выборки n.

Вероятность P выбирают, исходя из конкретных требований решаемой задачи. На практике обычно пользуются вероятностью $p=0,95$.

Сравним значения, приведенные в табл. 2 и 5 приложений. Например, в табл. 2 значению вероятности $p=0,95$ соответствует $t=1,96$. Проследим по таблице, как изменяется значение t с ростом n при $p=0,95$:

n	5	15	20	30	50	100	∞
$t_{n;\,0,95}$	2,78	2,15	2,093	2,045	2,009	1,984	1,960

Из анализа табличных данных видно, что уже при $n=50$ распределение Стьюдента мало отличается от нормального. Так, например, если $p=0,95$, то $z=1,96$ и $t_{50;\,0,95}=2,009$, т. е. $2,009-1,960=0,049$*. При малых значениях n t-распределение заметно отличается от нормального. Например, если $p=0,95$, то $z=1,96$ и $t_{5;\,0,95}=2,78$, т. е. $2,78-1,96=0,82$.

В заключение еще раз подчеркнем, что статистика $t=\dfrac{\bar{X}-\mu}{\hat{s}}\sqrt{n}$ имеет распределение Стьюдента, если случайная величина X подчиняется нормальному закону распределения, а средняя $\bar{X}$ вычисляется по выборочным данным $X_1, X_2, \ldots, X_n$, полученным в результате независимых наблюдений.

---

* В специальной литературе можно познакомиться с доказательством того, что при неограниченном возрастании объема выборки распределение Стьюдента стремится к нормальному закону распределения.

§ 5.8. Распределение дисперсии в выборках из нормальной генеральной совокупности. Распределение χ^2 Пирсона

Рассмотрим закон распределения выборочной дисперсии, рассчитанной для наблюдений, взятых из нормальной совокупности. Так как состав выборки подвержен случайности, то выборочную дисперсию, как и $\overline{X}$, следует рассматривать как случайную величину и говорить о законе распределения выборочной дисперсии. При анализе распределения дисперсии выборки следует иметь в виду следующие два случая: 1) математическое ожидание случайной величины известно; 2) математическое ожидание неизвестно.

Случай 1. Предположим, что математическое ожидание случайной величины известно. Условимся считать, что случайная величина X подчиняется нормальному закону распределения с параметрами μ, σ^2, а X_1, X_2, ..., X_n — ряд независимых наблюдений, каждое из которых подчиняется нормальному закону распределения с математическим ожиданием μ и дисперсией σ^2. Тогда выборочная дисперсия вычисляется по формуле

$$s_*^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2.$$

Разделим обе части этого равенства на σ^2 и умножим на n. Имеем

$$\frac{ns_*^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n} t_i^2.$$

В § 3.4 величина $(X_i-\mu)/\sigma$ обозначена t. Статистика t имеет нормальный закон распределения с параметрами $M(t)=0$ и $D(t)=1$.

В силу независимости значений X_i значения t_i также независимы, а это определяет независимость значений t_i^2 $(i=\overline{1, n})$.

Итак, пусть

$$\chi^2 = ns_*^2/\sigma^2 = \sum_{i=1}^{n} t_i^2.$$

Случайная величина, представляющая собой сумму квадратов независимых случайных величин $t_i (i=\overline{1, n})$, каждая из которых подчиняется нормальному закону распределения с параметрами ($\mu=0$, $\sigma^2=1$), называется *случайной величиной с распределением* χ^2 и $k=n$ степенями свободы. Дифференциальная функция распределения представлена на рис. 5.2.

Распределение статистики χ^2 не зависит ни от математического ожидания μ случайной величины X, ни от

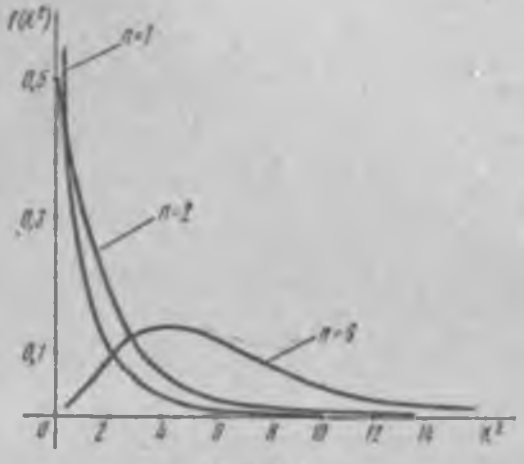

Рис. 5.2

дисперсии σ^2, а зависит лишь от объема выборки n. Найдем математическое ожидание распределения χ^2:

$$M(\chi^2) = M\left(\sum_{i=1}^{n} t_i^2\right) = \sum_{i=1}^{n} M(t_i^2) = \sum_{i=1}^{n} D(t) = n.$$

Следовательно, математическое ожидание случайной величины с распределением χ^2 и $k=n$ степенями свободы равно числу степеней свободы. В специальной литературе можно найти доказательство того, что дисперсия распределения χ^2 равна удвоенному числу степеней свободы, т. е. $D(\chi^2)=2k$. Дифференциальная функция распределения χ^2 сложна, и интегрирование ее является весьма трудоемким процессом, поэтому составлены таблицы распределения χ^2.

В приложениях приведена табл. 7 для вычисления вероятности того, что случайная величина, подчиняю-

щаяся закону χ^2 с известным числом степеней свободы, превысит некоторое фиксированное значение $\chi^2_{k;\alpha}$, т. е. $P(\chi^2 > \chi^2_{k;\alpha})$. Известно, что

$$P(\chi^2_{k;\alpha}) = \int_0^{\chi^2} f(\chi^2)\, d\chi^2.$$

Рассмотрим на конкретных примерах правило пользования таблицей.

Пример 5.3. Пусть для случайной величины, подчиняющейся закону χ^2 с числом степеней свободы $k=7$, требуется определить отклонение $\chi^2_{k;\alpha}$, вероятность превышения которого равна 0,05, т. е. найти $\chi^2_{k;\alpha}$, для которого $P(\chi^2 > \chi^2_{k;\alpha}) = 0{,}05$.

Решение. Искомое число $\chi^2_{k;\alpha}$ находится на пересечении столбца 0,05 и строки 7 табл. 7 приложений, т. е. $\chi^2_{7;0,05} = 14{,}1$, поэтому $P(\chi^2 > 14{,}1) = 0{,}05$.

Пример 5.4. Проведено 10 независимых наблюдений над случайной величиной X, распределенной нормально с математическим ожиданием $\mu=0$ и дисперсией $\sigma^2=4$. Найти вероятность того, что сумма квадратов результатов наблюдения превысит 16.

Решение. Пусть $x_1, x_2, \ldots, x_{10}$ — независимые наблюдения случайной величины X. Необходимо вычислить вероятность $P(\sum_{i=1}^{10} x_i^2 > 16)$. Так как по условию каждая из случайных величин x_i $(i=\overline{1,10})$ имеет нормальное распределение с $\mu=0$ и $\sigma^2=4$, то отношение $x_i/\sigma = x_i/2$ $(i=\overline{1,10})$ также подчиняется нормальному закону распределения с $\mu=0$ и $\sigma^2=1$, следовательно, сумма квадратов $\sum_{i=1}^{10}(x_i/2)^2$ имеет распределение χ^2 с $k=10$ степенями свободы. Поэтому

$$P\left(\sum_{i=1}^{10} x_i^2 > 16\right) = P\left[\sum_{i=1}^{10}\left(\frac{x_i}{2}\right)^2 > \frac{16}{4}\right] = P(\chi^2 > 4).$$

По табл. 7 приложений находим при $k=10$, т. е. в строке 10, число, близкое к 4. Таким числом является 3,94. Вероятность, соответствующая этому числу, находится в заголовке столбца: $P(\chi^2 > 4) \approx 0{,}95$.

Случай 2. Рассмотрим закон распределения выборочной дисперсии, когда математическое ожидание случайной величины X неизвестно. Как и прежде, считаем, что случайная величина X подчиняется нормальному закону распределения с параметрами μ, σ^2, а $X_1, X_2, \ldots, X_n$ есть ряд независимых наблюдений над случайной ве-

личиной X. Тогда дисперсия выборки вычисляется по формуле

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Примем без доказательства тот факт, что случайная величина $n\hat{s}^2/\sigma^2$ имеет распределение χ^2 с $k=n-1$ степенями свободы.

§ 5.9. Понятие доверительного интервала. Доверительная вероятность

В § 5.3 были рассмотрены некоторые выборочные характеристики, которые лучше всего в смысле несмещенности, эффективности и состоятельности оценивали параметры распределения генеральной совокупности. Так, например, если случайная величина распределена в генеральной совокупности по нормальному закону распределения, то лучшей оценкой математического ожидания по выборке является средняя арифметическая. Следовательно, математическое ожидание оценивалось одним числом — средней арифметической.

Оценку неизвестного параметра генеральной совокупности одним числом называют *точечной оценкой*. Наряду с точечным оцениванием статистическая теория оценивания параметров занимается вопросами интервального оценивания. Задачу интервального оценивания в самом общем виде можно сформулировать так: по данным выборки построить числовой интервал, относительно которого с заранее выбранной вероятностью можно сказать, что внутри этого интервала находится оцениваемый параметр. *Интервальное оценивание* особенно необходимо при малом числе наблюдений, когда точечная оценка мало надежна.

Доверительным интервалом $[\tilde{\Theta}_n^{(1)};\ \tilde{\Theta}_n^{(2)}]$ для параметра Θ называют такой интервал, относительно которого можно с заранее выбранной вероятностью $p=1-\alpha$, близкой к единице, утверждать, что он содержит неизвестное значение параметра Θ, т. е.

$$P\left[\tilde{\Theta}_n^{(1)} < \Theta < \tilde{\Theta}_n^{(2)}\right] = 1-\alpha.$$

Чем меньше для выбранной вероятности $\left|\tilde{\Theta}_n^{(2)}-\tilde{\Theta}_n^{(1)}\right|$, тем точнее оценка неизвестного параметра Θ, и, наоборот, если этот интервал велик, то оценка, произведенная с его помощью, мало пригодна для практики. Так как концы доверительного интервала $\tilde{\Theta}_n^{(1)}$ и $\tilde{\Theta}_n^{(2)}$ зависят от элементов выборки, т. е. ими определяются, то значения $\tilde{\Theta}_n^{(1)}$ и $\tilde{\Theta}_n^{(2)}$ могут меняться от выборки к выборке. Вероятность $p=1-\alpha$ принято называть *доверительной вероятностью,* а число α — *уровнем значимости.*

Остановимся на вопросе о том, какими принципами следует руководствоваться при выборе доверительной вероятности. Выбор доверительной вероятности не является математической задачей, а определяется конкретно решаемой проблемой. Для иллюстрации этого положения приведем следующий пример. Пусть на двух предприятиях вероятность выпуска годных изделий $p=1-\alpha=0{,}99$, т. е. вероятность выпуска бракованных изделий $\alpha=0{,}01$. Можно ли в рамках математической теории, т. е. не интересуясь характером выпускаемых изделий, решить вопрос о том, мала или велика вероятность α?

Пусть одно предприятие выпускает электролампы, а другое — парашюты. Если на 100 ламп встретится одна бракованная, то с этим можно мириться при условии, что выбросить 1% ламп дешевле, чем перестроить технологический процесс. Если же на 100 парашютов встретится один бракованный, что может повлечь за собой серьезные последствия, то с таким положением мириться никак нельзя. Следовательно, в первом случае вероятность брака α приемлема, а во втором — нет, поэтому выбор доверительной вероятности $p=1-\alpha$ следует производить, исходя из конкретных условий задачи.

§ 5.10. Построение доверительного интервала для математического ожидания при известном σ

Пусть случайная величина X распределена нормально, причем среднее квадратическое отклонение σ этого распределения известно. Требуется оценить неизвестное математическое ожидание μ. Наилучшей оценкой математического ожидания в смысле несмещенности, состоятельности и эффективности, как следует из § 5.3, явля-

ется выборочное среднее $\bar{X}$. Из § 5.7 известно, что выборочное среднее $\bar{X}$ распределено нормально с параметрами $M(\bar{X})=\mu$, $D(\bar{X})=\sigma^2/n$; нормированное отклонение $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ распределено также нормально с параметрами $\mu=0$, $\sigma^2=1$, поэтому вероятность любого отклонения $|\bar{X}-\mu|$ может быть вычислена по формуле

$$P\left(\left|\frac{\bar{X}-\mu}{\sigma}\sqrt{n}\right|<z_p\right)=\Phi(z). \qquad (5.6)$$

Задавая определенную доверительную вероятность $p=1-\alpha=\Phi(z)$, по табл. 2 приложений можно определить значение z_p. Для оценки μ преобразуем формулу (5.6):

$$P\left(|\bar{X}-\mu|<z_p\sigma/\sqrt{n}\right)=\Phi(z),$$

или

$$P\left(-z_p\sigma/\sqrt{n}<(\bar{X}-\mu)<z_p\sigma/\sqrt{n}\right)=\Phi(z),$$

откуда

$$P\left(\bar{X}-z_p\sigma/\sqrt{n}<\mu<\bar{X}+z_p\sigma/\sqrt{n}\right)=\Phi(z). \qquad (5.7)$$

Таким образом, с вероятностью (надежностью) $\Phi(z)=1-\alpha$ можно утверждать, что интервал

$$\left(\bar{X}-z_p\sigma/\sqrt{n};\quad \bar{X}+z_p\sigma/\sqrt{n}\right) \qquad (5.8)$$

является доверительным для оценки μ.

Замечание. Пусть $\Delta=z_p\sigma/\sqrt{n}$. Тогда: 1) при фиксированном значении z_p с возрастанием n величина Δ уменьшается, следовательно, точность интервального оценивания увеличивается; 2) увеличение надежности оценки $p=1-\alpha=\Phi(z)$ ведет к увеличению z_p, так как функция $\Phi(z)$ — возрастающая. При фиксированном объеме выборки n величина Δ возрастает с увеличением p, что ведет к уменьшению точности оценки.

Пример 5.5. Случайная величина X имеет нормальное распределение с известным средним квадратическим отклонением $\sigma=2$. Найти доверительный интервал для оценки неизвестного математического ожидания μ по выборочной средней $\bar{x}$, если объем выборки $n=16$ и задана доверительная вероятность $p=1-\alpha=0{,}95$.

Решение. По табл. 2 приложений находим значение z_p, соответствующее вероятности 0,95, имеем $z_{0{,}95}=1{,}96$. Далее находим

$$\Delta=\left(z_p\sigma/\sqrt{n}\right)=(1{,}96\cdot 2)/\sqrt{16}=(1{,}96\cdot 2)/4=0{,}98.$$

Следовательно, $(\bar{X}-0{,}98;\ \bar{X}+0{,}98)$ — доверительный интервал для оценки μ с вероятностью $p=0{,}95$. Например, если $\bar{x}=4{,}1$, то дове-

рительный интервал имеет следующие доверительные границы: $\bar{x}-0{,}98=4{,}1-0{,}98=3{,}12$, $\bar{x}+0{,}98=4{,}1+0{,}98=5{,}08$, поэтому можно записать, что $3{,}12<\mu<5{,}08$.

Так как $\bar{X}$ — случайная величина, возможные значения которой меняются от выборки к выборке, то концы доверительного интервала $(\bar{X}-\Delta;\ \bar{X}+\Delta)$ также являются случайными величинами, меняющимися от выборки к выборке.

Если выборки повторять и для каждой из них определять границы доверительного интервала, то при большом числе опытов частость интервалов, которые покроют неизвестное значение, мало отличается (по теореме Бернулли) от выбранной доверительной вероятности $1-\alpha$.

В рассмотренном примере выбрана доверительная вероятность $p=0{,}95$. Это значит, что если произведено достаточно большое количество выборок, то 95% из них определяют такие доверительные границы, внутри которых будет находиться оцениваемый параметр μ, и лишь 5% интервалов его не содержат. Поэтому было бы ошибкой истолковать равенство $P(3{,}12<\mu<5{,}08)=0{,}95$ как вероятность того, что математическое ожидание, заключенное в интервале [3,12; 5,08], равно 0,95. Математическое ожидание может либо попасть в этот интервал (и тогда вероятность события $[3{,}12<\mu<5{,}08]$ равна единице), либо не попасть (и тогда вероятность события $[3{,}12<\mu<5{,}08]$ равна нулю).

Следовательно, доверительную вероятность не следует связывать с оцениваемым параметром — она связана лишь с границами интервала, определяемыми случайностью выборки.

§ 5.11. Построение доверительного интервала для математического ожидания при неизвестном σ

Пусть случайная величина X распределена нормально, причем среднее квадратичное отклонение σ этого распределения неизвестно. Требуется оценить неизвестное математическое ожидание μ. Как следует из § 5.7, случайная величина $t=(\bar{X}-\mu)\sqrt{n}/\hat{s}$ распределена по закону Стьюдента, поэтому, выбрав вероятность $p=1-\alpha$ и зная объем выборки n, можно, пользуясь табл. 6 приложений, найти такое $t_{n,p}$, что

$$P\left(\left(|\bar{X}-\mu|\right)\sqrt{n}/\hat{s}<t_{n,p}\right)=1-\alpha. \qquad (5.9)$$

Произведем преобразование формулы (5.9), позволяющее оценить μ:

$$P\left(|\bar{X}-\mu|<t_{n,p}\hat{s}/\sqrt{n}\right)=1-\alpha,$$

или

$$P\left(-t_{n,p}\hat{s}/\sqrt{n}<(\bar{X}-\mu)<t_{n,p}\hat{s}/\sqrt{n}\right)=1-\alpha,$$

откуда

$$P\left(\bar{X} - t_{n,p}\hat{s}/\sqrt{n} < \mu < \bar{X} + t_{n,p}\hat{s}/\sqrt{n}\right) = 1 - \alpha.$$

Поэтому с вероятностью (надежностью) $p = 1 - \alpha$ можно утверждать, что интервал

$$\left[\bar{X} - t_{n,p}\hat{s}/\sqrt{n};\quad \bar{X} + t_{n,p}\hat{s}/\sqrt{n}\right] \qquad (5.10)$$

является доверительным для оценки μ. Несмотря на кажущееся сходство формул (5.8) и (5.10), между ними существует различие, заключающееся в том, что в формуле (5.10) коэффициент $t_{n,p}$ зависит не только от доверительной вероятности, но и от количества элементов в выборке. Особенно это различие заметно при малом количестве наблюдений.

Пример 5.6. Пусть требуется построить доверительный интервал для оценки неизвестного математического ожидания μ при $n = 9$, $p = 0{,}95$, $\bar{x} = 6$, $\hat{s} = 3$.

Р е ш е н и е. Построим сначала доверительный интервал по формуле (5.8), приняв $\sigma = \hat{s}$, а затем по формуле (5.10) и сравним полученные результаты.

По табл. 2 приложений значению $p = 0{,}95$ соответствует $z_{0,95} = 1{,}96$, следовательно,

$$\bar{x} - \sigma z_p/\sqrt{n} < \mu < \bar{x} + \sigma z_p/\sqrt{n}.$$

Подставим числовые данные:

$$6 - 3 \cdot 1{,}96/\sqrt{9} < \mu < 6 + 3 \cdot 1{,}96/\sqrt{9},$$

следовательно, $4{,}04 < \mu < 7{,}96$.

В табл. 5 приложений значениям $p = 0{,}95$; $n = 9$ соответствует $t_{9;\ 0,95} = 2{,}31$, поэтому

$$\bar{x} - \hat{s}t_{n,p}/\sqrt{n} < \mu < \bar{x} + \hat{s}t_{n,p}/\sqrt{n},$$

или $6 - 3 \cdot 2{,}31/\sqrt{9} < \mu < 6 + 3 \cdot 2{,}31/\sqrt{9}$. Окончательно имеем $3{,}69 < \mu < 8{,}31$.

В первом случае длина интервала равна $7{,}96 - 4{,}04 = 3{,}92$, во втором $8{,}31 - 3{,}69 = 4{\cdot}62$, т. е. по формуле (5.8) получен меньший доверительный интервал, чем по формуле (5.10). Объясним этот факт. Полученный при малом количестве наблюдений с помощью закона Стьюдента более широкий доверительный интервал для оценки математического ожидания по сравнению с результатом, полученным при использовании нормального закона распределения, вовсе не свидетельствует о недо-

статке распределения Стьюдента, а скорее наоборот — о его преимуществе, так как при использовании t-распределения учитывается, что никакой дополнительной информации о дисперсии генеральной совокупности σ^2, кроме той, которую дает выборка, нет. Поэтому использование при оценке математического ожидания нормального закона распределения при малом n и неизвестной дисперсии привело бы к неоправданному сужению доверительного интервала.

Итак, для оценки математического ожидания случайной величины, распределенной нормально, по малым выборкам при неизвестном σ следует пользоваться t-распределением Стьюдента.

§ 5.12. Построение доверительного интервала для дисперсии

Пусть случайная величина X распределена нормально. Требуется построить доверительный интервал для дисперсии генеральной совокупности σ^2 либо по выборочной дисперсии s_*^2, либо по $\hat{s}^2$.

Построение доверительного интервала для дисперсии основывается на том, что случайная величина ns_*^2/σ^2 имеет распределение χ^2 с $k=n$ степенями свободы, а величина $n\hat{s}^2/\sigma^2$ имеет распределение χ^2 с $k=n-1$ степенями свободы (см. теоремы, приведенные в § 5.8.

Подробно рассмотрим построение доверительного интервала для второго случая, так как именно он наиболее часто встречается на практике. Итак, для выбранной доверительной вероятности $p=1-\alpha$, учитывая, что $n\hat{s}^2/\sigma^2$ имеет распределение χ^2 с $k=n-1$ степенями свободы, можно записать

$$P\left(\chi_1^2 < n\hat{s}^2/\sigma^2 < \chi_2^2\right) = 1-\alpha.$$

Далее по таблице χ^2-распределения нужно выбрать такие два значения χ_1^2 и χ_2^2, чтобы площадь, заключенная под дифференциальной функцией распределения χ^2 между χ_1^2 и χ_2^2, была равна $1-\alpha$. Как видно из рис. 5.3, *а*, *б*, значения χ_1^2 и χ_2^2 можно варьировать так, чтобы значение заштрихованной площади остава-

лось постоянным ($s_1=s_2$) и равным $p=1-\alpha$ ($p=1-\alpha=s_1=s_2$). Обычно χ_1^2 и χ_2^2 выбирают такими, чтобы

$$P(\chi^2<\chi_1^2)=P(\chi^2>\chi_2^2)=\alpha/2,$$

т. е. площади, заштрихованные на рис. 5.4, равны между собой. Тогда имеем

$$P\left(\chi_1^2<n\hat{s}^2/\sigma^2<\chi_2^2\right)=1-\underbrace{P(\chi^2<\chi_1^2)}_{\alpha/2}-\underbrace{P(\chi^2>\chi_2^2)}_{\alpha/2}.$$

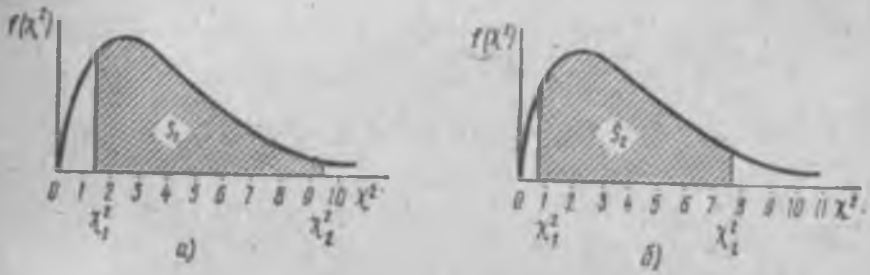

Рис. 5.3

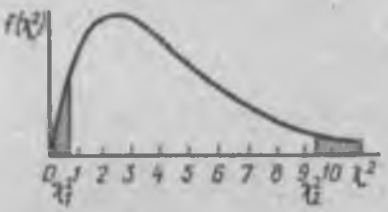

Рис. 5.4

Так как табл. 7 приложений содержит лишь $P(\chi^2>\chi_{k,\alpha}^2)$, то для вычисления по ней $P(\chi^2<\chi_1^2)$ запишем следующее тождество: $P(\chi^2<\chi_1^2)=1-P(\chi^2>\chi_1^2)=\alpha/2$. Поэтому

$$P\left(\chi_1^2<n\hat{s}^2/\sigma^2<\chi_2^2\right)=1-[1-P(\chi^2>\chi_1^2)]-P(\chi^2>\chi_2^2)=$$
$$=P(\chi^2>\chi_1^2)-P(\chi^2>\chi_2^2).$$

Итак,

$$P\left(\chi_1^2<n\hat{s}^2/\sigma^2<\chi_2^2\right)=P(\chi^2>\chi_1^2)-P(\chi^2>\chi_2^2)=1-\alpha;$$
$$P(\chi^2>\chi_1^2)=(1-\alpha)+P(\chi^2>\chi_2^2)=1-\alpha/2.$$

Преобразуем двойное неравенство:

$$\chi_1^2<n\hat{s}^2/\sigma^2<\chi_2^2. \qquad (5.11)$$

Запишем неравенство, обратное неравенству (5.11). Заметим, что в этом случае знаки неравенства изменятся на противоположные, т. е. $1/\chi_1^2 > \sigma^2/\left(n\hat{s}^2\right) > 1/\chi_2^2$, или $1/\chi_2^2 < \sigma^2/\left(n\hat{s}^2\right) < 1/\chi_1^2$. Умножая обе части неравенства на положительное число $n\hat{s}^2$, отличное от нуля, окончательно получаем

$$n\hat{s}^2/\chi_2^2 < \sigma^2 < n\hat{s}^2/\chi_1^2.$$

Неравенство (5.11) эквивалентно последнему неравенству, поэтому

$$P\left(\chi_1^2 < n\hat{s}^2/\sigma^2 < \chi_2^2\right) = P\left(n\hat{s}^2/\chi_2^2 < \sigma^2 < n\hat{s}^2/\chi_1^2\right) = 1-\alpha.$$

Пример 5.7. Построить доверительный интервал с вероятностью $p=0{,}96$ для дисперсии σ^2 случайной величины X, распределенной нормально, если $\hat{s}^2=10$, $n=20$.

Решение. Доверительная вероятность $p=1-\alpha=0{,}96$; $\alpha=0{,}04$. По табл. 7 приложений при $p_2=\alpha/2=0{,}04/2=0{,}02$ и $k=n-1=19$ степеням свободы находим значение $\chi_2^2=33{,}7$; вероятности $p_1=1-\alpha/2=1-1/2\cdot 4/100=0{,}98$ и $n_1-1=19$ степеням свободы соответствует значение $\chi_1^2=8{,}6$. Тогда доверительный интервал для σ^2 можно записать в следующем виде: $(20\cdot 10)/33{,}17<\sigma^2<(20\cdot 10)/8{,}6$, или $5{,}935<\sigma^2<23{,}256$.

Для интервальной оценки среднего квадратического отклонения служит равенство

$$P\left(\sqrt{n}\,\hat{s}/\chi_2 < \sigma < \sqrt{n}\,\hat{s}/\chi_1\right) = 1-\alpha.$$

Вычислим оценку среднего квадратического отклонения σ по данным примера 5.7: $\sqrt{5{,}935}<\sigma<\sqrt{23{,}256}$, или $2{,}43<\sigma<4{,}82$.

Глава 6

Проверка статистических гипотез

§ 6.1. Понятие статистической гипотезы. Общая постановка задачи проверки гипотез

В естествознании, технике, экономике часто для выяснения справедливости того или иного факта прибегают к высказыванию гипотез, которые можно проверить статистически, т. е. опираясь на результаты наблюдений в случайной выборке.

Под *статистической гипотезой* понимают всякое высказывание о генеральной совокупности, проверяемое по выборке. Статистические гипотезы классифицируют на гипотезы о законах распределения и гипотезы о параметрах распределения. Так, например, гипотеза о том, что производительность труда рабочих, выполняющих одинаковую работу в одинаковых организационно-технических условиях, имеет нормальный закон распределения, является гипотезой о законе распределения. Гипотеза о том, что средние размеры деталей, производимые на однотипных, параллельно работающих станках, не различаются между собой, является гипотезой о параметрах распределения.

Сформулируем в общем виде задачу статистической проверки гипотезы о параметре распределения. Пусть $f(X, \theta)$ — закон распределения случайной величины X, зависящей от одного параметра θ. Предположим, что необходимо проверить гипотезу о том, что $\theta=\theta_0$. Назовем эту гипотезу *нулевой* и обозначим ее H_0. Гипотезу о том, что $\theta=\theta_1$, назовем *конкурирующей* и обозначим ее H_1. Заметим, что иногда гипотезу H_1 называют *альтернативной гипотезой* или *альтернативой*. Таким образом, перед нами стоит задача проверки гипотезы H_0 относительно конкурирующей гипотезы H_1 на основании

выборки, состоящей из n независимых наблюдений X_1, X_2, ..., X_n над случайной величиной X. Следовательно, все возможное множество выборок объема n можно разделить на два непересекающихся подмножества (обозначим их O и W) таких, что проверяемая гипотеза H_0 должна быть отвергнута, если наблюдаемая выборка попадет в подмножество W, и принята, если выборка принадлежит подмножеству O.

Подмножество W называют *критической областью*; O—*областью допустимых значений*. Так как подмножество O состоит из всех тех выборок объема n, которые не вошли в подмножество W, то подмножество W однозначно определяет подмножество O и наоборот, т. е. необходимо определить одно подмножество, второе же получается автоматически единственным образом. Возникает вопрос: какими принципами следует руководствоваться при построении критической области W? Эти принципы были сформулированы в работах известных математиков Е. Неймана и Э. Пирсона. При выборе критической области следует иметь в виду, что, принимая или отклоняя гипотезу H_0, можно допустить ошибки двух видов.

Ошибка первого рода состоит в том, что нулевая гипотеза H_0 отвергается, т. е. принимается гипотеза H_1, в то время как в действительности все же верна гипотеза H_0.

Ошибка второго рода состоит в том, что гипотеза H_0 принимается, в то время как верна гипотеза H_1.

Рассмотренные случаи наглядно иллюстрирует табл. 6.1.

Таблица 6.1

Гипотеза H_0	Верна	Неверна
Отвергается	Ошибка первого рода	Правильное решение
Принимается	Правильное решение	Ошибка второго рода

Для любой заданной критической области W условимся обозначать через α вероятность ошибки первого рода, через β — вероятность ошибки второго рода. Сле-

довательно, можно сказать, что при большом количестве выборок доля ложных заключений равна α, если верна гипотеза H_0, и β, если верна гипотеза H_1. В теории статистической проверки гипотез доказывается, что при фиксированном объеме выборки соответствующий выбор критической области W позволяет сделать как угодно малой либо α, либо β.

§ 6.2. Проверка гипотезы о равенстве центров распределения двух нормальных генеральных совокупностей при известном σ

Проверка гипотез о равенстве двух центров распределения имеет важное практическое значение. Действительно, иногда оказывается, что средний результат одной серии экспериментов заметно отличается от среднего результата другой серии. При этом возникает вопрос: можно ли объяснить обнаруженное расхождение средних случайными ошибками эксперимента или оно вызвано какими-либо незамеченными или даже неизвестными закономерностями? В промышленности задача сравнения средних часто возникает при выборочном контроле качества изделий, изготовленных на разных установках или при разных технологических режимах.

Сформулируем задачу сравнения двух центров распределения в общем виде. Рассмотрим две случайные величины X и Y, каждая из которых подчиняется нормальному закону распределения. Пусть имеются две независимые выборки n_1 и n_2 из генеральных совокупностей X и Y. Необходимо проверить нулевую гипотезу H_0, заключающуюся в том, что $M(X)=M(Y)$, относительно альтернативной гипотезы H_1, состоящей в том, что $|M(X)-M(Y)|>0$.

В настоящем параграфе рассматривается случай, когда генеральные дисперсии σ_X^2 и σ_Y^2 известны. Так как о значениях математических ожиданий $M(X)$ и $M(Y)$ ничего не известно, то для проверки гипотезы H_0 используем их наилучшие оценки $\overline{X}$ и $\overline{Y}$. Как известно (см. § 5.7), средние $\overline{X}$ и $\overline{Y}$ имеют нормальный закон распределения с параметрами $[M(X); \sigma_X^2/n_1]$ и $[M(Y); \sigma_Y^2/n_2]$. Выборки независимы, поэтому $\overline{X}$ и $\overline{Y}$ также независимы и случайная величина, равная разности между $\overline{X}$ и $\overline{Y}$, имеет нормальное распределение, причем

$$D(\bar{X}-\bar{Y}) = D(\bar{X}) + D(\bar{Y}) = \sigma_x^2/n_1 + \sigma_y^2/n_2.$$

Если гипотеза H_0 справедлива, то $M(\bar{X}-\bar{Y}) = M(\bar{X}) - M(Y) = 0$, следовательно, нормированная разность

$$z = (\bar{X}-\bar{Y})/\sqrt{\sigma_x^2/n_1 + \sigma_y^2/n_2}$$

подчиняется нормальному закону распределения с математическим ожиданием, равным нулю, и дисперсией, равной единице.

Выбирая, сообразуясь с условиями задачи, вероятность $p=1-\alpha$, можно по табл. 2 приложений определить статистику z_p, которая разделит множество z на два непересекающихся подмножества: область допустимых значений z и критическую область z. Те значения z, при которых $|z| \leqslant z_p$, образуют область допустимых значений; значения z, для которых $|z| > z_p$, определяют критическую область. Итак, для вероятности $p=1-\alpha$ критическая область определяется неравенством

$$|\bar{X}-\bar{Y}|/\sqrt{\sigma_x^2/n_1 + \sigma_y^2/n_2} > z_p.$$

Если $\sigma_x^2 = \sigma_y^2 = \sigma^2$, то критическая область определяется неравенством

$$|\bar{X}-\bar{Y}|/(\sigma\sqrt{1/n_1 + 1/n_2}) > z_p.$$

Вероятность α отвечает событиям, которые при данных условиях исследования считаются практически невозможными. Чем меньше α, тем меньше вероятность отклонить проверяемую гипотезу, если она верна (совершить ошибку первого рода); с уменьшением уровня значимости α увеличивается риск принять проверяемую гипотезу, когда она неверна (увеличивается вероятность ошибки второго рода).

При проверке гипотезы о равенстве центров распределения двух нормальных генеральных совокупностей при заданном уровне значимости α контролируется лишь ошибка первого рода, но нельзя сделать вывод о степени риска, связанного с принятием неверной гипотезы, т. е. с возможностью совершения ошибки второго рода.

Пример 6.1. В результате двух серий измерений с количеством измерений $n_1=25$ и $n_2=50$ получены следующие средние значения исследуемой величины: $\bar{x}=9{,}79$ и $\bar{y}=9{,}60$. Можно ли с надежностью $p=0{,}99$ объяснить это расхождение случайными причинами, если известно, что средние квадратические отклонения в обеих сериях измерений $\sigma_x=\sigma_y=0{,}30$?

Решение. Вычислим нормированную разность:

$$|z| = \frac{|\bar{x} - \bar{y}|}{\sigma\sqrt{1/n_1 + 1/n_2}} = \frac{|7{,}79 - 9{,}60|}{0{,}30\sqrt{1/25 + 1/50}} = 2{,}59.$$

Сравним значение $|z| = 2{,}59$ с критическим значением $z_{0.99} = 2{,}576$. Имеем $2{,}59 > 2{,}576$, поэтому с надежностью 0,99 можно считать расхождение средних неслучайным (значимым), так как попадание в область значений $|z| > 2{,}576$ при нашей гипотезе практически невозможно. Заметим, однако, что для значений $|z| < 2{,}576$ еще нельзя утверждать, что гипотеза подтвердилась; можно только признать допустимость гипотезы для рассмотренных выборочных наблюдений до тех пор, пока более обстоятельные исследования не позволят сделать противоположное заключение.

Следовательно, с помощью проверки статистических гипотез можно лишь отвергнуть проверяемую гипотезу, но никогда нельзя доказать ее справедливость.

§ 6.3. Проверка гипотезы о равенстве центров распределения нормальных генеральных совокупностей при неизвестном σ. Метод исключения грубых ошибок в наблюдениях

Рассмотрим две случайные величины X и Y, каждая из которых подчиняется нормальному закону распределения с математическими ожиданиями $M(X)$ и $M(Y)$. Условимся считать, что $\sigma_x^2 = \sigma_y^2 = \sigma^2$; числовое значение σ^2 неизвестно.

Пусть имеются две независимые выборки n_1 и n_2 соответственно из генеральных совокупностей X и Y. Необходимо проверить нулевую гипотезу H_0, состоящую в том, что $M(X) = M(Y)$ относительно альтернативной гипотезы H_1, заключающейся в том, что $|M(X) - M(Y)| > 0$.

Для оценки $M(X)$ и $M(Y)$ используем их наилучшие оценки по выборке $\bar{x}$ и $\bar{y}$, а для оценки σ^2 — выборочные оценки

$$\hat{s}_x^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x}) \text{ и } \hat{s}_y^2 = \frac{1}{n_2 - 1}(y_i - \bar{y})^2.$$

Так как генеральные совокупности X и Y имеют одинаковые дисперсии, то для оценки σ^2 целесообразно использовать результаты обеих выборок. В математической статистике доказано, что лучшей оценкой для σ^2 в данном случае является

$$\hat{s}^2 = \frac{\hat{s}_x^2 (n_1 - 1) + \hat{s}_y^2 (n_2 - 1)}{n_1 + n_2 - 2}.$$

Если гипотеза H_0 справедлива, то случайная величина $(\overline{X} - \overline{Y})$ подчиняется нормальному закону распределения с математическим ожиданием, равным нулю, и дисперсией $\sigma^2 (1/n_1 + 1/n_2)$. Действительно,

$$M(\overline{X} - \overline{Y}) = M(\overline{X}) - M(\overline{Y}) = 0,$$

$$D(\overline{X} - \overline{Y}) = D(\overline{X}) + D(\overline{Y}) = \sigma^2/n_1 + \sigma^2/n_2 = \sigma^2 (1/n_1 + 1/n_2).$$

Величина σ^2 неизвестна, поэтому возникает вопрос: какой статистикой оценить $D(\overline{X} - \overline{Y})$? В качестве выборочной оценки $D(\overline{X} - \overline{Y})$ обычно принимают оценку $\hat{s}^2_{(\overline{X} - \overline{Y})} = (1/n_1 + 1/n_2) \hat{s}^2$, так как

$$M\left[(1/n_1 + 1/n_2) \hat{s}^2\right] = (1/n_1 + 1/n_2) M\left(\hat{s}^2\right) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times$$

$$\times M\left[\frac{\hat{s}_X^2 (n_1 - 1) + \hat{s}_Y^2 (n_2 - 1)}{n_1 + n_2 - 2}\right] = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times$$

$$\times \frac{(n_1 - 1) M\left(\hat{s}_X^2\right) + (n_2 - 1) M\left(\hat{s}_Y^2\right)}{n_1 + n_2 - 2} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times$$

$$\times \frac{(n_1 - 1) \sigma^2 + (n_2 - 1) \sigma^2}{n_1 + n_2 - 2} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sigma^2.$$

Известно (см. § 5.7), что если случайная величина $(\overline{X} - \overline{Y})$ подчиняется нормальному закону распределения, то статистика

$$t = \frac{(\overline{X} - \overline{Y}) - M(\overline{X} - \overline{Y})}{\hat{s}_{(\overline{X} - \overline{Y})}} =$$

$$= \frac{(\overline{X} - \overline{Y}) - M(\overline{X} - \overline{Y})}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1 - 1) \hat{s}_X^2 + (n_2 - 1) \hat{s}_Y^2}{n_1 + n_2 - 2}}}$$

имеет t-распределение Стьюдента, причем $k = n_1 + n_2 - 2$. Число k называют *числом степеней свободы* t-распределения. (Это понятие было введено в § 5.8.)

Если гипотеза H_0 справедлива, то статистику t можно записать в виде

$$t = \frac{(\bar{X} - \bar{Y})}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1 - 1)\,\hat{s}_X^2 + (n_2 - 1)\,\hat{s}_Y^2}{n_1 + n_2 - 2}}}.$$

Выбрав вероятность $p = 1 - \alpha$, по таблице t-распределения можно определить критическое значение $t_{n_1+n_2-2;\alpha}$, для которого $P(|t| > t_{n_1+n_2-2;\,\alpha}) = \alpha$. Если вычисленное значение $|t| > t_{n_1+n_2-2;\,\alpha}$, то с надежностью $p = 1 - \alpha$ можно считать расхождение средних значимым (неслучайным).

Пример 6.2. Пусть при тех же условиях, что и в примере 6.1 ($n_1 = 25$; $n_2 = 50$; $\bar{x} = 9{,}79$; $\bar{y} = 9{,}60$), средние квадратические отклонения в обеих генеральных совокупностях неизвестны (но равны). В этом случае для проверки гипотезы H_0 необходимо подсчитать $\hat{s}_x^2$ и s_y^2. Пусть $\hat{s}_x^2 = 0{,}28$; $\hat{s}_y^2 = 0{,}33$. Далее имеем:

$$\hat{s} = \sqrt{(0{,}28^2 \cdot 24 + 0{,}33^2 \cdot 49)/73} \approx 0{,}31;$$

$$t = (\bar{x} - \bar{y}) \Big/ \left(\hat{s}\sqrt{1/n_1 + 1/n_2}\right) = 0{,}19 \Big/ \left(0{,}31\sqrt{0{,}04 + 0{,}02}\right) = 2{,}47.$$

Вероятности $p = 0{,}99$ и числу степеней свободы $k = 73$ в таблице t-распределения соответствует $t_{73;\,0{,}01} = 2{,}65$.

Так как $2{,}47 < 2{,}65$, то с надежностью 0,99 нельзя считать расхождение средних значимым.

Рассмотренный выше критерий дает возможность сравнивать центры распределения при допущении о равенстве дисперсий σ_x^2 и σ_y^2 в генеральных совокупностях.

Более интересна общая постановка задачи, которая давала бы возможность сравнивать центры распределения без ограничений, накладываемых на дисперсии. Формулировка задачи в общем виде получила название *проблемы Беренса — Фишера*, которая пока еще не решена.

Приведенные ранее критерии позволяют сравнивать два параметра, однако, последовательно сравнивая, можно делать заключения о соотношении между любым конечным числом центров распределения. Только что рассмотренный критерий можно применять для исключения грубых ошибок при проведении наблюдений.

Грубые ошибки возникают из-за ошибок в отсчетах показаний измерительных приборов; в вычислениях при

измерении — из-за неправильного использования средств измерения, невнимательности оператора и т. д. Результаты наблюдений, проведенные с такими ошибками, как правило, резко отличаются от среднего результата данной серии наблюдений.

Наиболее надежным методом исключения грубых ошибок является браковка подозрительного результата в момент получения аномального наблюдения, когда можно установить причину его появления. Если этот момент упущен или условия эксперимента не позволяют решить подобную задачу, то исследователю приходится в дальнейшем решать вопрос о принадлежности резко выделяющегося результата наблюдений (выброса) к остальным наблюдениям. Пусть X_0, X_1, X_2, ..., X_n — совокупность результатов наблюдений, причем наблюдение X_0 резко выделяется. Для ряда наблюдений X_1, X_2, ..., X_n рассчитывают $\bar{X}$ и $\hat{s}$. При справедливости нулевой гипотезы (о принадлежности X_0 к остальным наблюдениям) статистика $t=(X_0-\bar{X})/\hat{s}$ имеет t-распределение Стьюдента с числом степеней свободы $k=n-1$.

Альтернативная гипотеза в этом случае запишется так: $H_1: X_0>\bar{X}$. Если в качестве альтернативной гипотезы проверяется $H_1: X_0<\bar{X}$, то статистика $t=(\bar{X}-X_0)/s$.

Проверка нулевой гипотезы производится аналогично тому, как это было сделано в случае сравнения двух центров распределения. Однако в этом случае для определения $t_{\text{табл}}$ необходимо использовать табл. 6 приложений, специально предназначенную для проверки принадлежности резко выделяющихся значений рассматриваемой совокупности, либо табл. 5 приложений, определяя $t_{\text{табл}}$ по $k=n-1$ и $p=1-2\alpha$. В этом случае критерий t называется *односторонним* критерием t в отличие от *двустороннего* критерия, рассматриваемого в § 6.3.

Пример 6.3. Рассматривают ряд наблюдений

x_0	x_1	x_2	x_3	x_4	x_5	x_6
25,5	24,1	23,8	23,5	23,4	23,3	23,2

Является ли значение $x_0=25{,}5$ резко выделяющимся при $p=0{,}95$?

Решение. Для проверки гипотезы H_0 (о принадлежности x_0 к остальным наблюдениям) необходимо найти:

$$\bar{x}=(24{,}1+23{,}8+23{,}5+23{,}4+23{,}3+23{,}2)/6\approx 23{,}6;$$

$$\hat{s}^2=1/5\,[(24{,}1-23{,}6)^2+(23{,}8-23{,}6)^2+(23{,}5-23{,}6)^2+$$
$$+(23{,}4-23{,}6)^2+(23{,}3-23{,}6)^2+(23{,}2-23{,}6)^2]\approx 0{,}12;$$

$$\hat{s} = \sqrt{0,12} \approx 0,35;$$

$$t = (25,5 - 23,6)/0,35 = 1,9/0,35 \approx 5,43;$$

$t_{6;\ 0,95} = 3,04$; $5,43 > 3,04$. Согласно t-критерию, наблюдение x_0 следует отбросить.

§ 6.4. F-распределение и проверка гипотезы о равенстве дисперсий двух нормальных генеральных совокупностей

Гипотезы о дисперсиях имеют особенно большое значение в технике, так как измеряемая дисперсией величина рассеяния характеризует такие исключительно важные показатели, как точность машин, приборов, технологических процессов и т. д.

Сформулируем гипотезу о равенстве дисперсий двух нормальных генеральных совокупностей. Рассмотрим две случайные величины X и Y, каждая из которых подчиняется нормальному закону распределения с дисперсиями σ_X^2 и σ_Y^2. Пусть из генеральных совокупностей X и Y извлечены две независимые выборки объемами n_1 и n_2. Проверим гипотезу H_0 о том, что $\sigma_X^2 = \sigma_Y^2$ относительно альтернативной гипотезы H_1, заключающейся в том, что $\sigma_X^2 > \sigma_Y^2$. Для оценки σ_X^2 используется выборочная дисперсия $\hat{s}_X^2$, а для оценки σ_Y^2 — выборочная дисперсия $\hat{s}_Y^2$, следовательно, задача проверки гипотезы H_0 сводится к сравнению дисперсий $\hat{s}_X^2$ и $\hat{s}_Y^2$.

Для построения критической области с выбранной надежностью выводов необходимо исследовать совместный закон распределения оценок $\hat{s}_X^2$ и $\hat{s}_Y^2$. Таким совместным законом распределения статистик $\hat{s}_X^2$ и $\hat{s}_Y^2$ является F-распределение, называемое *распределением Фишера — Снедекора.*

Рассмотрим случайную величину X, подчиняющуюся нормальному закону распределения с параметрами (μ, σ^2). Произведем из генеральной совокупности X две независимые выборки n_1 и n_2. Для оценки σ^2 можно использовать выборочные дисперсии $\hat{s}_1^2$ и $\hat{s}_2^2$, рассчитанные по элементам первой и второй выборок. Как следует из

§ 5.8, случайные величины $\hat{s}{}_1^2 n_1/\sigma^2$ и $\hat{s}{}_2^2 n_2/\sigma^2$ распределены по закону χ^2 с $k_1=n_1-1$ и $k_2=n_2-1$ степенями свободы. Случайную величину F, определяемую отношением

$$F=\frac{n_2}{n_1}\,\frac{\hat{s}{}_1^2\, n_1/\sigma^2}{\hat{s}{}_2^2\, n_2/\sigma^2}=\frac{\hat{s}{}_1^2}{\hat{s}{}_2^2},$$

называют *случайной величиной с распределением Фишера — Снедекора*. Заметим, что всегда можно так ввести обозначения, что окажется $\hat{s}{}_1^2 \geqslant \hat{s}{}_2^2$, поэтому случайная величина F принимает значения, не меньшие единицы.

Известна формула дифференциальной функции распределения случайной величины F. Оказывается, дифференциальный закон распределения случайной величины F не содержит неизвестных параметров (μ, σ^2) и их оценок $(\overline{X}, \hat{s}{}^2)$, опредляющих распределение X, а зависит лишь от числа наблюдений в выборках n_1 и n_2. Этот факт позволяет составить таблицы распределения случайной величины F, в которых различным значениям уровня значимости и различным сочетаниям величин k_1 и k_2 соответствуют такие значения $F(\alpha, k_1, k_2)$, для которых справедливо равенство $P[F>F_{(\alpha, k_1, k_2)}]=\alpha$.

Теперь вернемся к задаче проверки гипотезы H_0 о равенстве дисперсий. Вычислив выборочные дисперсии $\hat{s}{}_1^2$ и $\hat{s}{}_2^2$, найдем их отношение $F=\hat{s}{}_1^2/\hat{s}{}_2^2$, причем в числителе запишем большую из двух вычисленных оценок дисперсии σ^2 (это условие соблюдается для уменьшения объема таблиц). Затем, выбрав необходимый уровень значимости α, по таблице F-распределения находим число $F_{(\alpha, k_1, k_2)}$, которое сравниваем с вычисленным F. Если окажется, что $F>F_{(\alpha, k_1, k_2)}$, то проверяемая гипотеза H_0 отвергается, если $F\leqslant F_{(\alpha, k_1, k_2)}$, то выборочные наблюдения не противоречат проверяемой гипотезе.

Пример 6.4. На двух токарных станках обрабатывают втулки. Отобраны две пробы: из втулок, сделанных на первом станке, $n_1=10$ шт., на втором станке — $n_2=15$ шт. По данным этих выборок рассчитаны выборочные дисперсии $\hat{s}{}_1^2=9{,}6$ (для первого станка) и

$\hat{s}_2^2=5,7$ (для второго станка). Следовательно, $F=\hat{s}_1^2/\hat{s}_2^2=9,6/5,7=$ $=1,68$ с числом степеней свободы $k_1=10-1=9$ и $k_2=15-1=14$. Проверить при уровне значимости $\alpha=0,05$ гипотезу о том, что станки обладают различной точностью. (Размеры втулок подчиняются нормальному закону распределения с одним и тем же математическим ожиданием.)

Решение. Чтобы проверить гипотезу о равенстве дисперсий $\sigma_1^2=\sigma_2^2$ с уровнем значимости $\alpha=0,05$ при $k_1=9$ и $k_2=14$, по таблице F-распределения находим $F_{0,05;\,9;\,14}=2,65$. Итак, имеем $1,68<$ $<2,65$, следовательно, предположение о равенстве дисперсий не противоречит наблюдениям, иными словами, пока нет оснований считать, что станки обладают различной точностью.

Отметим, что при проверке гипотезы о равенстве дисперсий при заданном уровне значимости α контролируется лишь ошибка первого рода, но нельзя сказать о степени риска, связанного с принятием неверной гипотезы, т. е. с возможностью ошибки второго рода.

§ 6.5. Проверка гипотез о законе распределения. Критерий согласия χ^2

В предыдущих параграфах настоящей главы рассматривались гипотезы, относящиеся к отдельным параметрам распределения случайной величины, причем закон ее распределения предполагался известным. Однако во многих практических задачах точный закон распределения исследуемой случайной величины неизвестен, т. е. является гипотезой, которая требует статистической проверки.

Обозначим через X исследуемую случайную величину. Пусть требуется проверить гипотезу H_0 о том, что эта случайная величина подчиняется закону распределения $F(x)$. Для проверки гипотезы произведем выборку, состоящую из n независимых наблюдений над случайной величиной X. По выборке можно построить эмпирическое распределение $F^*(x)$ исследуемой случайной величины. Сравнение эмпирического $F^*(x)$ и теоретического распределений производится с помощью специально подобранной случайной величины — критерия согласия. Существует несколько критериев согласия: χ^2 Пирсона, Колмогорова, Смирнова и др.

Критерий согласия χ^2 Пирсона — наиболее часто употребляемый критерий для проверки гипотезы о законе распределения.

Рассмотрим этот критерий. Разобьем всю область изменения X на l интервалов Δ_1, Δ_2, ..., Δ_l и подсчитаем количество элементов n_i, попавших в каждый из интервалов Δ_i. Предполагая известным теоретический закон распределения $F(x)$, всегда можно определить p_i (вероятность попадания случайной величины X в интервал Δ_i), тогда теоретическое число значений случайной величины X, попавших в интервал Δ_i, можно рассчитать по формуле np_i. Результаты проведенных расчетов объединим в табл. 6.2.

Таблица 6.2

Интервалы Δ_i	Δ_1	Δ_2	...	Δ_i	...	Δ_l
Эмпирические частоты m_i	m_1	m_2	...	m_i	...	m_l
Теоретические частоты np_i	np_1	np_2	...	np_i	...	np_l

Здесь $m_1+m_2+...+m_i...+m_l=n$; $p_1+p_2+...+p_l=1$.

Если эмпирические частоты сильно отличаются от теоретических, то проверяемую гипотезу H_0 следует отвергнуть, в противоположном случае — принять.

Сформулируем критерий, который бы характеризовал степень расхождения между эмпирическими и теоретическими частотами. В литературе по математической статистике доказывается, что статистика

$$\chi^2 = \sum_{i=1}^{l} \frac{(m_i - np_i)^2}{np_i}$$

имеет распределение χ^2 с $k=l-r-1$ степенями свободы. Здесь r — число параметров распределения $F(x)$, рассчитанных по выборке.

Правило применения критерия χ^2 сводится к следующему. Рассчитав значение χ^2 и выбрав уровень значимости критерия α, по таблице χ^2-распределения определяют $\chi^2_{k;\alpha}$. Если $\chi^2 > \chi^2_{k;\alpha}$, то гипотезу H_0 отвергают, если $\chi^2 \leqslant \chi^2_{k;\alpha}$, то гипотезу принимают. Очевидно, что при проверке гипотезы о законе распределения контролируется лишь ошибка первого рода.

В заключение заметим, что необходимым условием применения критерия Пирсона является наличие в каждом из интервалов по меньшей мере 5—10 наблюдений. Если количество наблюдений в отдельных интервалах очень мало (порядка 1—2), то имеет смысл объединить некоторые интервалы. Проиллюстрируем применение критерия χ^2 на двух примерах. В примере 6.5 разберем случай, когда ставится гипотеза о том, что исследуемая случайная величина подчиняется закону Пуассона, а в примере 6.6 — нормальному закону распределения.

Пример 6.5. На телефонной станции производились наблюдения за числом X неправильных соединений в минуту. Наблюдения в течение часа дали следующие результаты: 3; 1; 3; 1; 4; 2; 2; 4; 0; 3; 0; 2; 2; 0; 2; 1; 4; 3; 3; 1; 4; 2; 2; 1; 1; 2; 1; 0; 3; 4; 1; 3; 2; 7; 2; 0; 0; 1; 3; 3; 1; 2; 4; 2; 0; 2; 3; 1; 2; 5; 1; 1; 0; 1; 1; 2; 2; 1; 1; 5. Определить среднюю арифметическую и дисперсию неправильных соединений в минуту и проверить выполнение основного условия для распределения Пуассона $[M(X)=\sigma^2]$. Найти теоретическое распределение Пуассона и проверить степень согласия теоретического и эмпирического распределений по критерию χ^2 Пирсона при уровне значимости $\alpha=0{,}05$.

Решение. Упорядочим результаты наблюдений, записав их в следующую таблицу:

x_i	0	1	2	3	4	5	6	7
m_i	8	17	16	10	6	2	0	1

Обозначим через $\bar{x}$ среднее число неправильных соединений в минуту. Имеем

$$\bar{x} = (0\cdot 8 + 1\cdot 17 + 2\cdot 16 + 3\cdot 10 + 4\cdot 6 + 5\cdot 2 + 6\cdot 0 + 7\cdot 1)/60 = 2.$$

Вычислим выборочную оценку дисперсии:

$$\hat{s}^2 = {}^1/_{60}\,[8\cdot(0-2)^2 + 17\cdot(1-2)^2 + 16\cdot(2-2)^2 + (3-2)^2\cdot 10 + 6\cdot(4-2)^2 + 2\cdot(5-2)^2 + 0\cdot(6-2)^2 + 1\cdot(7-2)^2] \approx 2{,}1.$$

Необходимое условие для распределения Пуассона $\bar{x}=\hat{s}^2=2$ практически выполняется. Запишем теоретический закон Пуассона в виде

$$P(m) = \frac{2^m}{m!}\,e^{-2}.$$

Так как значение математического ожидания неизвестно, то подставим вместо него оценку $\bar{x}$ по выборке. Имеем:

$$P(0) = \frac{2^0}{0!} e^{-2} = 0,1353 = p_0^*, \quad P(1) = \frac{2^1}{1!} e^{-2} = 0,2707 = p_1,$$

$$P(2) = \frac{2^2}{2!} e^{-2} = 0,2707 = p_2, \quad P(3) = \frac{2^3}{3!} e^{-2} = 0,1804 = p_3,$$

$$P(4) = \frac{2^4}{4!} e^{-2} = 0,0902 = p_4, \quad P(5) = \frac{2^5}{5!} e^{-2} = 0,0361 = p_5,$$

$$P(6) = \frac{2^6}{6!} e^{-2} = 0,0120 = p_6, \quad P(7) = \frac{2^7}{7!} e^{-2} = 0,0034 = p_7.$$

Запишем полученные результаты в табл. 6.3.

Таблица 6.3

i	0	1	2	3	4	5	6	7
x_i	0	1	2	3	4	5	6	7
m_i	8	17	16	10	6	2	0	1
np_i	8,1	16,24	16,24	10,62	5,41	2,166	0,72	0,204

Объединим столбцы 5, 6 и 7, так как в каждом из них мало наблюдений. Тогда таблица преобразуется (табл. 6.4).

Таблица 6.4

i	0	1	2	3	4	5
x_i	0	1	2	3	4	5 и более
m_i	8	17	16	10	6	3
np_i	8,1	16,24	16,24	10,82	5,41	3,09

Вычислим значение χ^2:

$$\chi^2 = (8-8,1)^2/8,1 + (17-16,24)^2/16,24 + (16-16,24)^2/16,24 + (10-10,82)^2/10,82 + (6-5,41)^2/5,41 + (3-3,09)^2/3,09 \approx 0,2.$$

* Напомним, что $0! = 1$.

Примем уровень значимости $\alpha=0{,}05$. Количество интервалов $l=6$. По выборке вычислен один параметр, которым определяется закон Пуассона, — математическое ожидание; следовательно, $r=1$. Поэтому число степеней свободы $k=l-r-1=6-1-1=4$. В табл. 7 приложений значениям $\alpha=0{,}05$ и $k=4$ соответствует $\chi^2_{4;\,0{,}5}=9{,}5$. Имеем $9{,}5>0{,}2$, следовательно, нет оснований отвергнуть проверяемую гипотезу.

Таблица 6.5

№ интервала	x_i-x_{i+1}	m_i
1	40—41	20
2	41—42	112
3	42—43	154
4	43—44	73
5	44—45	13
6	45—46	2
Σ		374

Пример 6.6. Для управления качеством выплавляемой стали необходимо знание законов распределения ее механических свойств. Для исследования взято $n=374$ наблюдений над сталью марки 35 ГС. Необходимо проверить гипотезу о том, что механическое свойство (предел текучести) этой стали имеет нормальный закон распределения. По исходным наблюдениям рассчитаны: $\bar{x}=42{,}37$ кг/мм²; $\overset{\Delta}{s}=0{,}94$ кг/мм²; построен интервальный вариационный ряд (табл. 6.5). Уровень значимости α принять равным 0,01.

Решение. Как следует из анализа интервального вариационного ряда, необходимо объединить интервалы 5 и 6, так как в интервале 6 количество наблюдений менее 5. В результате объединения получим следующий ряд распределения:

x_i-x_{i+1}	40—41	41—42	42—43	43—44	44—46
m_i	20	112	154	73	15

Теперь найдем вероятность p_i $(i=\overline{1,5})$; p_1 выражает вероятность того, что случайная величина X, имеющая нормальный закон распределения, принимает значение, принадлежащее интервалу (40; 41), т. е.

$$p_1=P(40<X<41)\,P=\left(\frac{40-42{,}37}{0{,}94}<\frac{X-\mu}{\sigma}<\frac{41-42{,}37}{0{,}94}\right)=$$

$$=\tfrac{1}{2}\Phi\left(\frac{41-42{,}37}{0{,}94}\right)-\tfrac{1}{2}\Phi\left(\frac{40-42{,}37}{0{,}94}\right)=\tfrac{1}{2}\Phi(-1{,}46)-$$

$$-\tfrac{1}{2}\Phi(-2{,}52)=\tfrac{1}{2}\Phi(2{,}52)-\tfrac{1}{2}\Phi(1{,}46)=$$

$$=0{,}4940-0{,}4280=0{,}0660.$$

Аналогично получаем: $p_2=0{,}2762$; $p_3=0{,}3971$; $p_4=0{,}2128$; $p_5=0{,}0417$.

Для нахождения статистики χ^2 удобно составить таблицу (табл. 6.6).

Таблица 6.6

№ п/п	x_i+x_{i+1}	m_i	p_i	np_i	$\lvert m_i-np_i\rvert$	$(m_i-np_i)^2$	$\frac{(m_i-np_i)^2}{np_i}$
1	40—41	20	0,0660	24,76	4,76	21,56	0,871
2	41—42	112	0,2762	103,70	8,30	68,89	0,664
3	42—43	154	0,3971	148,52	5,48	29,03	0,127
4	43—44	73	0,2128	79,59	6,59	43,43	0,546
5	44—46	15	0,0417	15,60	0,60	0,36	0,023
Σ		374	0,9938	372,17			$2,231=\chi^2$

Количество интервалов $l=5$. По выборке рассчитаны два параметра: оценки для математического ожидания и дисперсии, следовательно, $r=2$. Число степеней свободы $k=l-r-1=5-2-1=2$.

В табл. 7 приложений $\alpha=0,01$ и $k=2$ соответствует $\chi^2_{2;0,01}=9,2$. Имеем $9,2>2,23$, следовательно, нет оснований отвергнуть гипотезу о нормальности распределения предела текучести.

Глава 7

Выборочный метод

§ 7.1. Статистическая теория выборочного метода

В гл. 5 и 6 отмечалось, что теория статистического оценивания параметров и проверки гипотез позволяет делать научно обоснованные выводы о всей генеральной совокупности по выборке из нее. Но при рассмотрении этих вопросов нигде не говорилось о том, как составлять выборку, чтобы она давала надежное представление о всей генеральной совокупности. Естественно, что не всякая выборка позволяет получить действительное представление о генеральной совокупности. Приведем пример. Автомат изготовляет шарики. Контролируется диаметр шарика. Шарики упакованы в ящик в два ряда: в нижнем ряду находятся шарики с диаметром 1,7 см$\leqslant D\leqslant$2 см, в верхнем — 1,5 см$\leqslant$ $\leqslant D<$1,7 см. Пусть требуется определить средний размер шарика в партии по выборке. Поскольку выборка может быть произведена из шариков, расположенных в верхнем ряду, то очевидно, что она не даст объективного представления о всей партии.

Чтобы по данным выборки можно было с уверенностью судить об интересующем признаке генеральной совокупности, выборка должна быть *представительной (репрезентативной)*. Выборка является репрезентативной, если она образована случайно. Существуют различные способы отбора. В зависимости от способа отбора различают выборки следующих типов: 1) случайная выборка с возвратом; 2) случайная выборка без возврата; 3) механическая; 4) типическая; 5) серийная.

Рассмотрим образование случайных выборок с возвратом и без возврата. Если выборку производят из

массы изделий (например, из ящика), то после тщательного перемешивания следует брать объекты случайно, т. е. так, чтобы все они имели одинаковую вероятность попасть в выборку. Часто для образования случайной выборки элементы генеральной совокупности предварительно нумеруют, а каждый номер записывают на отдельную карточку. В результате получают пачку карточек, число которых совпадает с объемом генеральной совокупности. После тщательного перемешивания из этой пачки берут по одной карточке. Объект, имеющий одинаковый номер с карточкой, считается попавшим в выборку. При этом возможны два принципиально различных способа образования выборочной совокупности.

Первый способ — вынутая карточка после фиксации ее номера возвращается в пачку, после чего карточки снова тщательно перемешиваются. Повторяя такие выборки по одной карточке, можно образовать выборочную совокупность любого объема. Выборочную совокупность, образованную по такой схеме, называют *случайной выборкой с возвратом* (повторная выборка).

Второй способ — вынутая карточка после записи ее номера обратно не возвращается. Повторяя по такой схеме выборки по одной карточке, можно получить выборочную совокупность любого заданного объема. Выборочную совокупность, образованную по этому способу, называют *случайной выборкой без возврата* (бесповторная выборка). Случайная выборка без возврата образуется в случае, если из тщательно перемешанной пачки сразу берут нужное число карточек.

Однако при большом объеме генеральной совокупности описанный выше способ образования случайной выборки с возвратом и без возврата оказывается очень трудоемким. В этом случае пользуются таблицами случайных чисел, в которых числа расположены в случайном порядке. Для того чтобы отобрать, например, 50 объектов из пронумерованной генеральной совокупности, открывают любую страницу таблицы случайных чисел и выписывают подряд 50 чисел; в выборку попадают те объекты, номера которых совпадают с выписанными случайными числами. Если случайное число таблицы окажется больше объема генеральной совокупности, то такое число пропускают.

Заметим, что различие между случайными выборками с возвратом и без возврата стирается, если они со-

ставляют незначительную часть большой генеральной совокупности.

При *механическом* способе образования выборочной совокупности подлежащие обследованию элементы генеральной совокупности отбирают через определенный интервал. Так, например, если выборка должна составлять 50% генеральной совокупности, то отбирают каждый второй элемент генеральной совокупности. Если выборка 10%-ная, то отбирают каждый 10-й ее элемент и т. д.

Следует отметить, что иногда механический отбор может не обеспечить репрезентативности выборки. Например, если отбирается каждый 20-й обтачиваемый валик, причем сразу же после отбора заменяют резец, то отобранными окажутся все валики, обточенные затупленными резцами. В таком случае необходимо устранить совпадение ритма отбора с ритмом замены резца, для чего следует отбирать хотя бы каждый 10-й валик из 20 обточенных.

При большом количестве выпускаемой однородной продукции, когда в ее изготовлении принимают участие различные станки и даже цехи, для образования репрезентативной выборки пользуются *типическим способом* отбора. В этом случае генеральную совокупность предварительно разбивают на непересекающиеся группы. Затем из каждой группы по схеме случайной выборки с возвратом или без возврата отбирают определенное число элементов. Они и образуют выборочную совокупность, которая называется *типической*.

Пусть, например, выборочным путем исследуется продукция цеха, в котором имеются 10 станков, производящих одну и ту же продукцию. Пользуясь схемой случайной выборки с возвратом или без возврата, отбирают изделия сначала из продукции, сделанной на первом, затем втором станках и т. д. Такой способ отбора позволяет образовать типическую выборку.

Иногда на практике бывает целесообразно использовать *серийный способ* отбора, идея которого заключается в том, что генеральную совокупность разбивают на некоторое количество непересекающихся серий и по схеме случайной выборки с возвратом или без возврата контролируют все элементы лишь отобранных серий. Например, если изделия изготовляются большой группой станков-автоматов, то сплошному обследованию подвер-

гают продукцию только нескольких станков. Серийным отбором пользуются в случае, если обследуемый признак колеблется в различных сериях незначительно.

Какому способу отбора следует отдать предпочтение в той или иной ситуации, следует судить, исходя из требований поставленной задачи и условий производства. Заметим, что на практике при составлении выборки часто используют одновременно несколько способов отбора в комплексе.

7.2. Оценка математического ожидания и дисперсии по случайной выборке с возвратом и без возврата

Проанализируем формулы, позволяющие оценить математическое ожидание и дисперсию по случайной выборке с возвратом и без возврата. Сначала рассмотрим образование повторной выборки. Элементы выборки X_1, X_2, ..., X_n можно рассматривать (см. § 5.1) как случайные величины, а способ образования выборки позволяет сделать заключение, что они независимы. Действительно, обозначим объем генеральной совокупности N. Тогда вероятность того, что X принимает конкретное значение x_1, равна $P(X=x_1)=1/N$; вероятность того, что X принимает значение x_2, также равна $P(X=x_2)=1/N$ и т. д. Следовательно, вероятность события $X=x_i$ не зависит от того, произошли ли события $X=x_i$. Поэтому на основании результатов, полученных в § 5.3, можно сделать вывод о том, что средняя $\bar{x}=(x_1+x_2+...+x_n)/n$, рассчитанная по элементам выборки, которая составлена по способу случайной выборки с возвратом, является наилучшей (несмещенной, состоятельной и в случае нормального распределения случайной величины в генеральной совокупности, эффективной) оценкой математического ожидания μ с дисперсией σ^2/n, где μ — математическое ожидание, а σ^2 — дисперсия генеральной совокупности. В том случае, когда значение математического ожидания неизвестно, для оценки генеральной дисперсии σ^2 по случайной выборке с возвратом пользуются состоятельной и несмещенной оценкой $\overset{\Delta}{s}{}^2$.

Пример 7.1. По схеме случайной выборки с возвратом обследован рост 625 мужчин. Определить с вероятностью 0,99 наибольшее отклонение среднего роста мужчин в выборке от среднего роста

мужчин в генеральной совокупности, если распределение роста мужчин в генеральной совокупности подчиняется нормальному закону распределения. Известно, что среднее квадратическое отклонение роста мужчин в генеральной совокупности равно 6 см.

Р е ш е н и е. Используя тот факт, что случайная величина (рост мужчин) подчиняется нормальному закону распределения, можно записать

$$P(|\bar{x}-\mu|<\Delta)=\Phi(\Delta/\sigma_{\bar{x}}).$$

Так как дисперсия средней арифметической в n раз меньше дисперсии случайной величины, то $\sigma^2_{\bar{x}}=\sigma^2/n$, или

$$\sigma_{\bar{x}}=\sigma/\sqrt{n}=6/\sqrt{625}=6/25=0{,}24.$$

Таким образом,

$$P(|\bar{x}-\mu|<\Delta)=\Phi(\Delta/0{,}24)=0{,}99.$$

По табл. 2 приложений находим $\Phi(z)=0{,}99$ при $z=2{,}58$. Следовательно, $\Delta/0{,}24=2{,}58$, откуда $\Delta=2{,}58\cdot 0{,}24=0{,}62$ см.

Итак, с вероятностью 0,99 можно утверждать, что средний рост мужчин в выборке отличается от среднего роста мужчин в генеральной совокупности не более чем на 0,62 см.

Если производится случайная выборка без возврата, то элементы $x_1, x_2,\ldots, x_n$ уже нельзя считать независимыми. Действительно, если обозначить объем генеральной совокупности N, то вероятность события $X_1=x_1$ равна $P(X=x_1)=1/N$, а вероятность события $X=x_2$ при условии, что элемент x_1 уже отобран, равна $P(x_2/x_1)=1/(N-1)$, т. е. вероятность события $X=x_2$ меняется в зависимости от того, имеет место событие $X=x_1$ или нет. Однако и в этом случае можно доказать, что средняя арифметическая является несмещенной и состоятельной оценкой математического ожидания, а ее дисперсия

$$D(\bar{X})=(\sigma^2/n)(N-n)/(N-1). \qquad (7.1)$$

Очевидно, если объем генеральной совокупности велик, а выборка представляет собой лишь ее небольшую часть, то выражение $(N-n)/(N-1)\to 1$ и результаты оценки дисперсии по случайной выборке с возвратом и без возврата мало различаются между собой.

Пример 7.2. Из 4000 электрических лампочек по схеме случайной выборки без возврата отобрано 200 лампочек, средняя продолжительность горения которых оказалась равной 1250 ч. Какова вероятность того, что средний срок службы лампочек, подчиняющийся нормальному закону распределения, находится в границах от 1220 до 1280 ч? Среднее квадратическое отклонение генеральной совокупности принять равным 150 ч.

Р е ш е н и е. Имеем: $N=4000$; $n=200$; $\bar{x}=1250$; $\sigma=150$. Нас интересует вероятность неравенства $1220<\mu<1280$. Рассмотрим ряд неравенств, эквивалентных записанному: $(1220-1250)<(\mu-\bar{x})<(1280-1250)$, или $-30<(\mu-\bar{x})<30$, или $|\mu-\bar{x}|<30$. Так как случайная величина подчиняется нормальному закону распределения, то

$$P(|\mu-\bar{x}|<30)=\Phi(30/\sigma_{\bar{x}}).$$

Теперь необходимо вычислить $\sigma_{\bar{x}}$, т. е. среднее квадратическое отклонение средней арифметической. Так как

$$\sigma_{\bar{x}}^2=\frac{\sigma^2}{n}\frac{N-n}{N-1},$$

то

$$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}=\frac{150}{\sqrt{200}}\sqrt{\frac{4000-200}{4000-1}}\approx 10{,}34.$$

Таким образом,

$$P(|\mu-\bar{x}|<30)\approx\Phi\left(\frac{30}{10{,}34}\right)=\Phi(2{,}90)=0{,}9962.$$

Если дисперсия генеральной совокупности неизвестна, то по случайной выборке без возврата ее оценивают по формуле

$$s^2=(N-1)\hat{s}^2/N,$$

где $\hat{s}^2=\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right)/(n-1)$.

Однако на практике дисперсию генеральной совокупности можно оценивать и в этом случае выборочной дисперсией $\hat{s}^2$, так как обычно объем генеральной совокупности велик и сомножитель $(N-1)/N$ можно принять равным единице.

§ 7.3. Вычисление объема выборки

После того как для оценки интересующего параметра генеральной совокупности обоснованно выбран способ образования выборки, рассчитывают необходимый объем выборки, задавшись желаемой степенью точности оценки Δ и доверительной вероятностью $(1-\alpha)$.

Рассмотрим задачу определения объема случайной выборки с возвратом и случайной выборки без возврата

для оценки математического ожидания по выборочной средней.

Сначала рассчитаем объем случайной выборки с возвратом. Если исследуемая случайная величина подчиняется нормальному закону распределения и ее среднее квадратическое отклонение известно, то

$$P\left(\bar{x}-z_p\frac{\sigma}{\sqrt{n}}<\mu<\bar{x}+z_p\frac{\sigma}{\sqrt{n}}\right)=\Phi(z_p).$$

Сообразуясь с требованиями конкретно решаемой задачи, выберем доверительную вероятность $1-\alpha=\Phi(z_p)$ и из табл. 2 приложений для заданной вероятности выпишем z_p. Длина доверительного интервала

$$\left[\left(\bar{x}+z_p\frac{\sigma}{\sqrt{n}}\right)-\left(\bar{x}-z_p\frac{\sigma}{\sqrt{n}}\right)\right]=2z_p\frac{\sigma}{\sqrt{n}}$$

определяет точность оцениваемого параметра. Если обозначить $z_p\sigma/\sqrt{n}=\Delta$, то по заданной точности Δ и найденному значению z_p всегда можно вычислить n. Действительно, $z_p^2\sigma^2/n=\Delta^2$, откуда

$$n=\left(z_p^2\sigma^2\right)/\Delta^2. \qquad (7.2)$$

Пример 7.3. Каким должен быть объем случайной выборки для оценки математического ожидания μ, если известно, что половина доверительного интервала $\Delta=0{,}5$, а доверительная вероятность $p=1-\alpha=0{,}95$? Дисперсия изучаемой случайной величины $\sigma^2=5$.

Решение. По табл. 2 приложений находим значение z_p, соответствующее $p=0{,}95$. Имеем $z_{0,95}=1{,}96$. Подставляя в формулу (7.2) исходные данные, получаем $n=(1{,}96^2\cdot 5)/0{,}5^2=76{,}8\approx 77$. Следовательно, для заданной доверительной вероятности и точности оценки объем выборки должен составлять не менее 77 наблюдений.

Если вычисленный объем n выборки слишком велик по сравнению с практически имеющимися возможностями, то его можно снизить либо уменьшив доверительную вероятность, либо увеличив доверительный интервал.

При $p=0{,}8064$ по табл. 2 приложений находим $z_{0,8064}=1{,}30$, тогда $n=(1{,}30^2\cdot 5)/0{,}5^2=33{,}8\approx 34$

Если среднее квадратическое отклонение σ случайной величины X, подчиняющейся нормальному закону распределения, неизвестно, то предварительно берут небольшую пробную выборку и по ее данным приближенно оценивают параметр σ^2. Эту оценку подставляют затем в формулу (7.2), которая в этом случае принимает вид

$$N=\left(t_{n,p}^2\hat{s}^2\right)/\Delta^2,$$

где $\hat{s}^2$ — выборочная дисперсия пробной выборки; $t_{n,p}$ — число, взятое из табл. 5 приложений, соответствующее вероятности p и числу наблюдений n.

Пусть теперь выборка, отбираемая для определения средней, является случайной выборкой без возврата. Если исследуемая случайная величина подчиняется нормальному закону распределения, то можно записать

$$P\left(\bar{X} - z_p \sigma_{\bar{X}} < \mu < \bar{X} + z_p \sigma_{\bar{X}}\right) = \Phi(z_p).$$

Как и прежде, сообразуясь с условиями конкретной задачи, выберем доверительную вероятность $1-\alpha=\Phi(z_p)$ и по табл. 2 приложений для заданной вероятности находим z_p.

Если дисперсия σ^2 генеральной совокупности известна, то

$$\sigma_{\bar{X}}^2 = (\sigma^2/n)(N-n)/(N-1),$$

и доверительный интервал для математического ожидания можно записать в виде

$$\bar{X} - z_p \sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}} < \mu < \bar{X} + z_p \sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}}.$$

В данном случае

$$\Delta^2 = z_p^2 \frac{\sigma^2}{n}\frac{N-n}{N-1},$$

откуда $n(N-1)\Delta^2 = z_p^2 \sigma^2 N - n\sigma^2 z_p^2$,

$$n = \frac{N z_p^2 \sigma^2}{(N-1)\Delta^2 + z_p^2 \sigma^2}. \qquad (7.3)$$

Очевидно, при $N \to \infty$ следует использовать формулу $n=(z_p^2\sigma^2)/\Delta^2$, так как

$$\lim_{N\to\infty} \frac{N z_p^2 \sigma^2}{(N-1)\Delta^2 + z_p^2 \sigma^2} = \lim_{N\to\infty} \frac{z_p^2 \sigma^2}{(N-1)\Delta^2/N + z_p^2 \sigma^2/N} = \frac{z_p^2 \sigma^2}{\Delta^2}.$$

Пример 7.4. Определить необходимый объем случайной выборки без возврата для определения средней продолжительности горения лампочек в партии из 5000 шт., чтобы с вероятностью 0,99 предельная ошибка выборки не превышала 25 ч. Генеральное среднее квадратическое отклонение принять равным 150 ч.

Р е ш е н и е. По формуле (7.3) при $N=5000$; $z_{0,99}=2,58$; $\sigma=150$; $\Delta=25$ найдем

$$n=\frac{5000\cdot 2,58\cdot 150^2}{4999\cdot 25^2+2,58^2\cdot 150^2}=228,7\approx 229.$$

Теперь решим вопрос: какая из двух выборок (случайная выборка с возвратом или без возврата) требует при заданном доверительном интервале 2Δ и доверительной вероятности $p=1-\alpha$ бо́льшего количества наблюдений в выборке? Обозначим через n' количество элементов случайной выборки без возврата. Таким образом, необходимо сравнить формулы

$$n'=\frac{Nz_p^2\sigma^2}{(N-1)\Delta^2+z_p^2\sigma^2} \text{ и } n=\frac{z_p^2\sigma^2}{\Delta^2}. \qquad (7.4)$$

Заметим, что $1/n=\Delta^2/(z_p^2\sigma^2)$. Разделим числитель и знаменатель формулы для n' на Δ^2:

$$n'=\frac{Nn}{N+(n-1)}. \qquad (7.5)$$

Как известно, значение дроби увеличится, если уменьшить ее знаменатель. Так как $n\geqslant 1$, то $n-1\geqslant 0$, поэтому дробь не уменьшится, если отбросить в знаменателе выражение $n-1$, следовательно, $n'\leqslant Nn/N=n$. Таким образом, $n'\leqslant n$, т. е. объем случайной выборки с возвратом не меньше объема случайной выборки без возврата.

Пример 7.5. Определить необходимый объем случайной выборки с возвратом и случайной выборки без возврата для оценки средней длины деталей в партии, состоящей из 800 деталей, если доверительная вероятность $p=0,92$, доверительный интервал $2\Delta=0,04$; дисперсия $\sigma^2=0,25$.

Р е ш е н и е. Имеем: $N=8000$; $z_{0,92}=1,75$; $\Delta=0,02$; $\sigma^2=0,25$. По формуле (7.4) находим $n=(1,75^2\cdot 0,25)/0,02^2=1914$. Значение n' вычисляем по формуле (7.5): $n'=(8000\cdot 1914)/(8000+1913)\approx 1545$. Итак, объем случайной выборки с возвратом больше, чем объем случайной выборки без возврата ($1914>1545$).

Глава 8

Основы дисперсионного анализа

§ 8.1. Общая идея дисперсионного анализа

Дисперсионный анализ — это статистический метод анализа результатов наблюдений, зависящих от различных, одновременно действующих факторов, выбор наиболее важных факторов и оценка их влияния. Дисперсионный анализ находит применение в различных областях науки и техники. Идея дисперсионного анализа, как и сам термин «дисперсия», принадлежат Р. Фишеру. Суть анализа заключается в разложении общей вариации случайной величины на независимые слагаемые, каждое из которых характеризует влияние того или иного фактора или их взаимодействия.

Факторами обычно называют внешние условия, влияющие на эксперимент. Это, например, температура и атмосферное давление, сила тяготения, тип оборудования и т. п. Нас интересуют факторы, действие которых значительно и поддается проверке. В условиях эксперимента факторы могут варьировать, благодаря чему можно исследовать влияние контролируемого фактора на эксперимент. В этом случае говорят, что фактор варьирует на разных уровнях или имеет несколько уровней. В зависимости от количества факторов, включенных в анализ, различают классификацию по одному признаку — *однофакторный анализ*, по двум признакам — *двухфакторный анализ* и многостороннюю классификацию — *перекрестную классификацию*, изучением которой занимается многофакторный анализ.

Для проведения дисперсионного анализа необходимо соблюдать следующие условия: результаты наблюдений должны быть независимыми случайными величинами, имеющими нормальное распределение и одинаковую дис-

персию. Только в этом случае можно оценить значимость полученных оценок дисперсий и математических ожиданий и построить доверительные интервалы.

§ 8.2. Однофакторный анализ

На практике возможен случай, когда на автоматической линии несколько станков параллельно выполняют некоторую операцию. Для правильного планирования последующей обработки важно знать, насколько однотипны средние размеры деталей, получаемые на параллельно работающих станках. Здесь имеет место лишь один фактор, влияющий на размер деталей, — станки, на которых они изготовляются. Исследователя интересует, насколько существенно влияние этого фактора на размеры деталей? Предположим, что совокупности размеров деталей, изготовленных на каждом станке, имеют нормальное распределение и равные дисперсии. Имеем m станков, следовательно, m совокупностей или уровней, на которых произведено n_1, n_2,..., n_m наблюдений. Для простоты рассуждений положим, что $n_1=n_2=...=n_m$. Размеры деталей, составляющие n_i наблюдений на i-м уровне, обозначим x_{i1}, x_{i2},..., x_{in}. Тогда все наблюдения можно представить в виде таблицы, которую назовем *матрицей наблюдений* (табл. 8.1).

Таблица 8.1

Уровни	Наблюдения			
	1	2	j	n
1	x_{11}	x_{12}	x_{1j}	x_{1n}
2	x_{21}	x_{22}	x_{2j}	x_{2n}
...	...	...	...	...
i	x_{i1}	x_{i2}	x_{ij}	x_{in}
...	...	...	...	...
m	x_{m1}	x_{m2}	x_{mj}	x_{mn}

Будем полагать, что для i-го уровня n наблюдений имеют среднюю β_i, равную сумме общей средней μ и вариации ее, обусловленной i-м уровнем фактора, т. е. $\beta_i=\mu+\gamma_i$. Тогда одно наблюдение можно представить в следующем виде:

$$x_{ij}=\mu+\gamma_i+\xi_{ij}=\beta_i+\xi_{ij}, \qquad (8.1)$$

где μ — общая средняя; γ — эффект, обусловленный i-м уровнем фактора; ξ_{ij} — вариация результатов внутри отдельного уровня. Член ξ_{ij} характеризует влияние всех не учтенных моделью (8.1) факторов. Согласно общей задаче дисперсионного анализа, нужно оценить существенность влияния фактора γ на размеры деталей. Общую вариацию переменной x_{ij} можно разложить на части, одна из которых характеризует влияние фактора γ, другая — влияние неучтенных факторов. Для этого необходимо найти оценку общей средней μ и оценки средних по уровням β_i. Очевидно, что оценкой β является средняя арифметическая n наблюдений i-го уровня, т. е. $\bar{x}_{i*} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$. Звездочка в индексе при x означает, что наблюдения фиксированы на i-м уровне. Средняя арифметическая всей совокупности наблюдений является оценкой общей средней μ, т. е.

$$\bar{x} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} \text{ или } \bar{x} = \frac{1}{m}\sum_{i=1}^{m}\bar{x}_{i*}.$$

Найдем сумму квадратов отклонений x_{ij} от $\bar{x}$, т. е. $\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2$. Представим ее в виде

$$Q = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i*}+\bar{x}_{i*}-\bar{x})^2 =$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i*})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{x}_{i*}-\bar{x})^2 +$$

$$+ 2\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i*})(\bar{x}_{i*}-\bar{x}), \qquad (8.2)$$

причем

$$S = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i*})(\bar{x}_{i*}-\bar{x}) = \sum_{j=1}^{n}(x_{ij}-\bar{x}_{i*})\sum_{i=1}^{m}(\bar{x}_{i*}-\bar{x}).$$

Но $\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i*})=0$, так как это есть сумма отклонений

переменных одной совокупности от средней арифметической этой же совокупности, т. е. $S=0$. Второй член суммы (8.2) запишем в виде

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{x}_{i*}-\bar{x})^2 = n\sum_{i=1}^{m}(\bar{x}_{i*}-\bar{x})^2.$$

Тогда основное тождество (8.2) можно представить следующим образом:

$$\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2}_{Q} = \underbrace{n\sum_{i=1}^{m}(\bar{x}_{i*}-\bar{x})^2}_{Q_1} + \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i*})^2}_{Q_2};$$

или

$$Q = Q_1 + Q_2. \qquad (8.3)$$

Слагаемое Q_1 является суммой квадратов разностей между средними уровней и средней всей совокупности наблюдений. Эта сумма называется *суммой квадратов отклонений между группами* и характеризует расхождение между уровнями. Величину Q_1 называют также *рассеиванием по факторам*, т. е. рассеиванием за счет исследуемого фактора. Слагаемое Q_2 является суммой квадратов разностей между отдельными наблюдениями и средней i-го уровня. Эта сумма называется *суммой квадратов отклонений внутри группы* и характеризует расхождение между наблюдениями i-го уровня. Величину Q_2 называют также *остаточным рассеиванием*, т. е. рассеиванием за счет неучтенных факторов. Наконец, Q называется *общей* или *полной суммой квадратов отклонений отдельных наблюдений от общей средней* $\bar{x}$.

Зная суммы квадратов Q, Q_1 и Q_2, можно оценить соответствующие дисперсии: общую, межгрупповую и внутригрупповую. Оценим дисперсии s_1^2, s_2^2, s^2:

$$s_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\bar{x}_{i*}-\bar{x})^2 = \frac{Q_1}{m-1},$$

$$s_2^2 = \frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i*})^2 = \frac{Q_2}{m(n-1)},$$

$$s^2 = \frac{1}{mn-1}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2.$$

Если влияние всех уровней фактора γ одинаково, то s_1^2 и s_2^2 — оценки общей дисперсии. Тогда для оценки существенности влияния фактора γ достаточно проверить гипотезу H_0: $s_1^2 = s_2^2$; для этого вычисляют статистику $F = s_1^2/s_2^2$ с $k_1 = m-1$ и $k_2 = m(n-1)$ степенями свободы. Затем по табл. 4 приложения для уровня значимости α находят критическое значение F_{α, k_1, k_2}. Если $F > F_{\alpha, k_1, k_2}$, то нулевая гипотеза отвергается и делается заключение о существенном влиянии фактора γ. При $F < F_{\alpha, k_1, k_2}$ нет оснований отвергать нулевую гипотезу и считают, что влияние фактора γ несущественно.

Сравнивая межгрупповую и остаточную дисперсии, по величине их отношения судят, насколько сильно проявляется влияние факторов.

Однофакторный дисперсионный анализ удобно представить в виде табл. 8.2.

Таблица 8.2

Компоненты дисперсии	Сумма квадратов	Число степеней свободы k	Средний квадрат	Оценка дисперсии
Межгрупповая	$\sum (\bar{x}_{i\cdot} - \bar{x})^2$	$m-1$	$\frac{1}{m-1} \times \sum_i (\bar{x}_{i\cdot} - \bar{x})^2$	s_1^2
Внутригрупповая	$\sum_{ij} (x_{ij} - \bar{x}_{i\cdot})^2$	$m(n-1)$	$\frac{1}{m(n-1)} \times \sum_{ij} (x_{ij} - \bar{x}_{i\cdot})^2$	s_2^2
Полная (общая)	$\sum_{ij} (x_{ij} - \bar{x})^2$	$mn-1$	$\frac{1}{mn-1} \times \sum_{ij} (x_{ij} - \bar{x})^2$	

Пример однофакторного анализа. Пусть имеется четыре партии сырья для текстильной промышленности. Из каждой партии отобрано по пять образцов и проведены испытания на определение величины разрывной нагрузки. Результаты испытаний приведены в табл. 8.3.

Таблица 8.3

Номер партии	Разрывная нагрузка				
1	200	140	170	145	165
2	190	150	210	150	150
3	230	190	200	190	200
4	150	170	150	170	180

Требуется выяснить, существенно ли влияние различных партий сырья на величину разрывной нагрузки.

Решение. В данном случае $m=4$, $n=5$. Среднюю арифметическую каждой строки вычисляем по формуле

$$\bar{x}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}.$$

Имеем: $\bar{x}_{1\bullet}=164$; $\bar{x}_{2\bullet}=170$; $\bar{x}_{3\bullet}=202$; $\bar{x}_{4\bullet}=164$.

Найдем среднюю арифметическую всех совокупностей:

$$\bar{x} = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}/(mn) = 3500/20 = 175.$$

Вычислим величины, необходимые для построения табл. 8.2:

сумму квадратов отклонений между группами Q_1 с $k_1=m-1$ степенями свободы

$$Q_1 = 5\sum_{i=1}^{4}(\bar{x}_{i\bullet}-\bar{x})^2 = 5\cdot 996; \quad k_1 = 4-1=3;$$

сумму квадратов отклонений внутри группы Q_2 с $k_2=mn-m$ степенями свободы

$$Q_2 = \sum_{i=1}^{4}\sum_{j=1}^{5}(x_{ij}-\bar{x}_{i\bullet})^2 = 7270; \quad k_2 = 20-4=16;$$

полную сумму квадратов Q с $k=mn-1$ степенями свободы

$$Q = \sum_{i=1}^{4}\sum_{j=1}^{5}(x_{ij}-\bar{x})^2 = 12\,250; \quad k = 20-1=19.$$

По найденным значениям составляем табл. 8.4.

Вычислим статистику $F=(4980\cdot 1/3)/(7270\cdot 1/16)=3{,}65$.

По табл. 4 приложений находим значение F при $k_2=16$ и $k_1=3$ степенях свободы и уровне значимости $\alpha=0{,}01$. Имеем $F_\alpha=9{,}01$. Вычисленное значение F меньше табличного, поэтому можно утверждать, что нулевая гипотеза не отвергается, а это значит, что различие между сырьем в партиях не влияет на величину разрывной нагрузки.

Таблица 84

Компоненты дисперсии	Суммы квадратов	Число степеней свободы	Средний квадрат
Межгрупповая	4 980	3	1660,0
Внутригрупповая	7 270	16	454,4
Полная	12 250	19	644,7

Следует осторожно подходить к истолкованию окончательных результатов, так как они предполагают нормальную плотность и тождественность дисперсий. Каждое из допущений требует проверки, основанной на тщательном анализе проведенных экспериментов.

§ 8.3. Двухфакторный анализ

Если на результативный признак влияют несколько факторов одновременно, то имеет место многофакторный анализ. Дисперсионный анализ в этом случае имеет свои особенности, так как необходимо учитывать взаимодействия между факторами.

Рассмотрим задачу оценки влияния двух одновременно действующих факторов. Предположим, что имеется несколько однотипных станков и несколько видов сырья. Требуется выяснить, значимо ли влияние различных станков и качество сырья в партиях на качество обрабатываемых деталей. Это типичная задача двухфакторного дисперсионного анализа.

Считаем, что предпосылки дисперсионного анализа выполнены. Пусть фактор A — влияние настройки станка, фактор B — влияние качества сырья. Имеем r станков, следовательно, r уровней фактора A, v партий сырья, следовательно, v уровней фактора B. Матрицу наблюдений можно представить в виде табл. 8.5. Пересечение i-го уровня фактора A с j-м уровнем фактора B образует ij-ю ячейку, в которую записывают наблюдения, полученные при одновременном исследовании факторов A и B на i-м и j-м уровнях.

Для простоты можно предположить, что имеем в ячейке только одно наблюдение x_{ij}. Предположим также, что между факторами A и B нет взаимодействия и что на i-м уровне фактора A наблюдения имеют сред-

Таблица 8.5

Партии сырья j / Станки i	B_1	B_2	...	B_j	...	B_v	$\bar{x}_{i\cdot}$
A_1	x_{11}	x_{12}	...	...	...	x_{1v}	$\bar{x}_{1\cdot}$
A_2	x_{21}	x_{22}	...	...	...	x_{2v}	$\bar{x}_{2\cdot}$
⋮	⋮	⋮	...	...	...	⋮	⋮
A_i				x_{ij}			
⋮							
A_r	x_{r1}	x_{r2}	...	...	...	x_{rv}	$\bar{x}_{r\cdot}$
$\bar{x}_{\cdot j}$	$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	...	...	...	$\bar{x}_{\cdot v}$	$\bar{x}$

нюю β_{iA}, а на j-м уровне фактора B наблюдения — среднюю β_{jB}. Тогда одно наблюдение можно представить в виде

$$x_{ij} = \mu + \gamma_i + g_j + e_{ij}, \qquad (8.4)$$

где μ — общая средняя; γ — эффект, обусловленный влиянием i-го уровня фактора A; g_j — эффект, обусловленный влиянием j-го уровня фактора B; e_{ij} — вариация результатов внутри отдельной ячейки (в случае одного наблюдения вариация равна нулю).

Оценками μ, β_{iA} и β_{jB} являются соответственно общая средняя $\bar{x} = \frac{1}{rv}\sum_{i=1}^{r}\sum_{j=1}^{v} x_{ij}$ и средние по уровням $\bar{x}_{i\cdot} = \frac{1}{v}\sum_{j=1}^{v} x_{ij}$, $\bar{x}_{\cdot j} = \frac{1}{r}\sum_{i=1}^{r} x_{ij}$.

Оценки общей дисперсии можно получить из основного тождества дисперсионного анализа. В двухфакторном дисперсионном анализе общая сумма квадратов от-

клонений от общей средней раскладывается согласно формуле (8.4) уже не на две, а на три части: часть общей суммы квадратов, обусловленную влиянием фактора A, часть, обусловленную влиянием фактора B, и часть, обусловленную влиянием неучтенных факторов. С помощью дисперсионных отношений можно выяснить, насколько существенно влияние каждой из этих частей. Действительно,

$$Q=\sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij}-\bar{x})^2=\sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij}-\bar{x}_{i*}-\bar{x}_{*j}+$$

$$+\bar{x}+\bar{x}_{i*}-\bar{x}+\bar{x}_{*j}-\bar{x})^2=\underbrace{v\sum_{i=1}^{r}(\bar{x}_{i*}-\bar{x})^2}_{Q_1}+\underbrace{r\sum_{j=1}^{v}(\bar{x}_{*j}-\bar{x})^2}_{Q_2}+$$

$$+\underbrace{\sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij}-\bar{x}_{i*}-\bar{x}_{*j}+\bar{x})^2}_{Q_3}=Q_1+Q_2+Q_3. \qquad (8.5)$$

Слагаемое Q_1 представляет собой сумму квадратов разностей между средними по строкам и общим средним и характеризует изменение признака по фактору A. Слагаемое Q_2 представляет собой сумму квадратов разностей между средними по столбцам и общим средним и характеризует изменение признака по фактору B. Слагаемое Q_3 называется остаточной суммой квадратов и характеризует влияние неучтенных факторов. Сумма Q называется общей или полной суммой квадратов отклонений отдельных наблюдений от общей средней. Оценим дисперсии:

$$s^2=\frac{1}{rv-1}\sum_{i=1}^{r}\sum_{j=1}^{v}(x_{ij}-\bar{x})^2=\frac{Q}{rv-1}, \qquad (8.6)$$

$$s_1^2=\frac{1}{r-1}v\sum_{i=1}^{r}(\bar{x}_{i*}-\bar{x})^2=\frac{Q_1}{r-1}, \qquad (8.7)$$

$$s_2^2=\frac{1}{v-1}r\sum_{j=1}^{v}(\bar{x}_{*j}-\bar{x})^2=\frac{Q_2}{v-1}, \qquad (8.8)$$

$$s_3^2 = \frac{1}{(r-1)(v-1)} \sum_{i=1}^{r} \sum_{j=1}^{v} (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}) =$$

$$= Q_3/[(r-1)(v-1)]. \qquad (8.9)$$

В двухфакторном анализе для выяснения значимости влияния факторов A и B на исследуемый признак сравнивают дисперсии по факторам с остаточной дисперсией. Вычисляют статистики $F_1 = s_1^2/s_3^2$ с $(r-1)$ и $(r-1)(v-1)$ степенями свободы и $F_2 = s_2^2/s_3^2$ с $(v-1)$ и $(r-1)(v-1)$ степенями свободы. Сравнение вычисленных статистик с табличными значениями и выводы о существенности влияния факторов производят так же, как и в однофакторном дисперсионном анализе.

Двухфакторный дисперсионный анализ удобно представить в виде табл. 8.6.

Таблица 8.6

Компонента дисперсии	Сумма квадратов	Число степеней свободы	Оценки дисперсий
Между средними по строкам	$Q_1 = v \sum_{i=1}^{r} (\bar{x}_{i\cdot} - \bar{x})^2$	$r-1$	s_1^2
Между средними по столбцам	$Q_2 = r \sum_{j=1}^{v} (\bar{x}_{\cdot j} - \bar{x})^2$	$v-1$	s_2^2
Остаточная	$Q_3 = \sum_{i=1}^{r} \sum_{j=1}^{v} (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2$	$(r-1)(v-1)$	s_3^2
Полная (общая)	$Q = \sum_{i=1}^{r} \sum_{j=1}^{v} (x_{ij} - \bar{x})^2$	$rv-1$	s^2

Рассмотрим на примере построение двухфакторного комплекса с одним наблюдением в ячейке. Имеем три уровня фактора B: B_1, B_2 и B_3 и два уровня фактора A: A_1 и A_2 (табл. 8.7). Для данного комплекса $r=2$, $v=3$, $n=r \cdot v=6$. В нижней строке табл. 8.7 и в правом крайнем столбце приведены средние значения по строкам и столбцам, т. е. по уровням факторов. Так, среднее по уровню фактора B_1 равно $\bar{x}_{\cdot 1}=(1+5)/2=3$; среднее по

Таблица 8.7

A \ B	B_1	B_2	B_3	$\bar{x}_{i\bullet}$
A_1	1	2	3	2
A_2	5	6	10	7
$\bar{x}_{\bullet j}$	3	4	6,5	4,5

уровню фактора A_1 равно $\bar{x}_{1\bullet}=(1+2+3)/3=2$. Общее среднее $\bar{x}=4{,}5$. По формуле (8.5) вычислим суммы квадратов:

$$Q_1 = 37{,}5;\ Q_2 = 13;\ Q_3 = 3;\ Q = 53{,}5.$$

Для нахождения оценок дисперсий воспользуемся формулами (8.6)—(8.9). Имеем:

$$s_1^2 = 37{,}5/1 = 37{,}5;\ s_2^2 = 13/2 = 6{,}5;\ s_3^2 = 3/1\cdot 2 = 1{,}5.$$

Теперь табл. 8.6 можно записать в следующем виде (табл. 8.8).

Таблица 8.8

Компонента дисперсий	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Между средними по строкам (фактор A)	37,5	1	37,5
Между средними по столбцам (фактор B)	13,0	2	6,5
Остаточная	3,0	2	1,5
Полная	53,5	5	10,7

Вычисляем F_A и F_B:

$$F_A = s_1^2/s_3^2 = 37{,}5/1{,}5 = 25;\ F_B = s_2^2/s_3^2 = 6{,}5/1{,}5 = 4{,}3.$$

Для уровня значимости $\alpha=0{,}05$ и $k_2=2$, $k_1=1$ степеней свободы по табл. 4 приложений находим значения F_α:

$$F_{\alpha(A)} = 18{,}51,\ F_A = 25;\quad F_{\alpha(B)} = 19{,}0,\ F_B = 4{,}3.$$

Сравнивая табличные значения с вычисленными, имеем: $F_A > F_{\alpha\,(A)}$; $F_B < F_{\alpha\,(B)}$. Полученные результаты позволяют сделать следующие выводы: нулевая гипотеза о равенстве средних по строкам не подтверждается, т. е. влияние фактора A на исследуемый признак значимо: нулевая гипотеза о равенстве средних по столбцам не опровергается, т. е. влияние фактора B на исследуемый признак незначимо.

Мы рассмотрели частный случай дисперсионного анализа при классификации по двум признакам: в ячейке одно наблюдение, взаимодействие между факторами отсутствует. В общем случае в ячейке может быть несколько наблюдений (как равное количество, так и неравное), между факторами может иметь место взаимодействие. Лучше, когда в ячейке равное количество наблюдений, так как при этом упрощаются вычисления.

Для общего случая двухфакторного анализа одно наблюдение можно представить в виде

$$x_{ijk} = \mu + \gamma_i + g_j + \nu_{ij} + e_{ijk},$$

где μ — общее среднее; γ_i — эффект, обусловленный влиянием i-го уровня фактора A; g_j — эффект, обусловленный влиянием j-го уровня фактора B; ν_{ij} — эффект взаимодействия факторов A и B; e_{ijk} — вариация внутри ячейки.

Основное тождество двухфакторного дисперсионного анализа с одинаковым количеством наблюдений в ячейке (n) имеет вид

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{v}\sum_{k=1}^{n}(x_{ijk} - \bar{x})^2 = nv\sum_{i=1}^{r}(\bar{x}_{i**} - \bar{x})^2 +$$

$$+ nr\sum_{j=1}^{n}(\bar{x}_{*j*} - \bar{x})^2 + n\sum_{i=1}^{r}\sum_{j=1}^{v}(\bar{x}_{ij*} - \bar{x}_{i**} - \bar{x}_{*j*} + \bar{x})^2 +$$

$$+ \sum_{i=1}^{r}\sum_{j=1}^{v}\sum_{k=1}^{n}(x_{ijk} - \bar{x}_{ij*})^2 = Q_1 + Q_2 + Q_3 + Q_4.$$

Здесь Q_1 и Q_2 имеют то же значение, что и в формуле (8.5); Q_3 — сумма квадратов, оценивающая взаимодействие факторов A и B; Q_4 — сумма квадратов, оценивающая вариацию внутри ячейки.

Матрицу наблюдений можно представить в виде табл. 8.9.

Таблица 8.9

A \ B	B_1	B_2	...	Bv	$\bar{x}_{i\cdot\cdot}$
A_1	$\bar{x}_{11\cdot}$ $x_{111}, x_{112}, \ldots, x_{11n}$	$\bar{x}_{12\cdot}$ $x_{121}, x_{122}, \ldots, x_{12n}$	...	$\bar{x}_{1v\cdot}$ $x_{1v1}, x_{1v2}, \ldots, x_{1vn}$	$\bar{x}_{1\cdot\cdot}$
A_2	$\bar{x}_{21\cdot}$ $x_{211}, x_{212}, \ldots, x_{21n}$	$\bar{x}_{22\cdot}$ $x_{221}, x_{222}, \ldots, x_{22n}$	...	$\bar{x}_{2v\cdot}$ $x_{2v1}, x_{2v2}, \ldots, x_{2vn}$	$\bar{x}_{2\cdot\cdot}$
...	...	...	$\bar{x}_{ij\cdot}$ $x_{ij1}, x_{ij2}, \ldots, x_{ijn}$	...	$\bar{x}_{i\cdot\cdot}$
A_r	$\bar{x}_{r1\cdot}$ $x_{r11}, x_{r12}, \ldots, x_{r1n}$	$\bar{x}_{r2\cdot}$ $x_{r21}, x_{r22}, \ldots, x_{r2n}$	...	$\bar{x}_{rv\cdot}$ $x_{rv1}, x_{rv2}, \ldots, x_{rvn}$	$\bar{x}_{r\cdot\cdot}$
$\bar{x}_{\cdot j\cdot}$	$\bar{x}_{\cdot 1\cdot}$	$\bar{x}_{\cdot 2\cdot}$	...	$\bar{x}_{\cdot v\cdot}$	$\bar{x}$

Здесь $x_{111}, x_{112}, \ldots, x_{rvn}$ — наблюдавшиеся значения исследуемого признака;

$$\bar{x}_{ij\cdot} = \frac{1}{n}\sum_{k=1}^{n} x_{ijk} \text{ — среднее значение в ячейке;}$$

$$\bar{x}_{i\cdot\cdot} = \frac{1}{v}\sum_{j=1}^{v} \bar{x}_{ij\cdot} \text{ — среднее значение по строке;}$$

$$\bar{x}_{\cdot j\cdot} = \frac{1}{r}\sum_{i=1}^{r} \bar{x}_{ij\cdot} \text{ — среднее значение по столбцу;}$$

$$\bar{x} = \frac{1}{rv}\sum_{i=1}^{r}\sum_{j=1}^{v} \bar{x}_{ij\cdot} \text{ — общее среднее;} \qquad (8.10)$$

r — число уровней фактора A; v — число уровней фактора B.

Порядок проведения дисперсионного анализа в этом случае такой же, как и прежде: сначала вычисляют суммы квадратов, оценки дисперсий, затем отношение дисперсий F сравнивают с табличным.

Схема анализа и порядок вычисления сумм приведены в табл. 8.10. Как видно из табл. 8.10, в схеме анализа

Таблица 8.10

Компонента дисперсий	Суммы квадратов	Число степеней свободы	Оценка дисперсии
Между средними по строкам (по фактору A)	$Q_1 = vn\sum_{i=1}^{r}(\bar{x}_{i\cdot\cdot} - \bar{x})^2$	$v-1$	s_1^2
Между средними по столбцам (по фактору B)	$Q_2 = rn\sum_{j=1}^{v}(\bar{x}_{\cdot j\cdot} - \bar{x})^2$	$r-1$	s_2^2
Взаимодействие	$Q_3 = n\sum_{i=1}^{r}\sum_{j=1}^{v}(\bar{x}_{ij\cdot} - \bar{x}_{i\cdot\cdot} - \bar{x}_{\cdot j\cdot} + \bar{x})^2$	$(v-1)(r-1)$	s_3^2
Остаточная	$Q_4 = \sum_{i=1}^{r}\sum_{j=1}^{v}\sum_{k=1}^{n}(x_{ijk} - \bar{x}_{ij\cdot})^2$	$rv(n-1)$	s_4^2
Полная сумма квадратов	$Q = \sum_{i=1}^{r}\sum_{j=1}^{v}\sum_{k=1}^{n}(x_{ijk} - \bar{x})^2$	$rvn-1$	s^2

появляется новая сумма квадратов Q_4 и несколько меняется структура суммы Q_3 (вместо x_{ijk} берется $\bar{x}_{ij\cdot}$). Появление суммы Q_4 обусловлено наличием нескольких наблюдений в ячейке. В предыдущей схеме (см. табл. 8.6) эта сумма отсутствовала, так как при одном наблюдении в ячейке разность $(x_{ijk} - \bar{x}_{ij\cdot})$ равна нулю. Сумма Q_4 характеризует влияние прочих случайных факторов (кроме факторов A, B и их взаимодействия), поэтому для определения значимости влияния факторов A и B величину дисперсии, обусловленную влиянием этих факторов, сравнивают с дисперсией, обусловленной влиянием прочих факторов. При этом вычисляют следующие отношения дисперсий: $F_A = s_1^2/s_4^2$, $F_B = s_2^2/s_4^2$, $F_{AB} = s_3^2/s_4^2$. Вычисленные значения сравнивают с табличными значениями F_α, которые получены для заданного уровня значимости и соответствующего числа степеней свободы.

Рассмотрим пример построения двухфакторного комплекса по приведенной схеме. В текстильной промышленности важным является выявление факторов, влияю-

щих на качество пряжи, с тем чтобы в дальнейшем их было можно регулировать. В табл. 8.11 приведены данные о величине разрывной нагрузки в зависимости от наладки машины и партии сырья.

Таблица 8.11

Уровень наладки машины \ Партии сырья	B_1					B_2				
A_1	190	260	170	170	170	190	150	210	150	150
A_2	150	250	220	140	180	230	190	200	190	200
A_3	190	185	135	195	195	150	170	150	170	180

При каждом уровне наладки машины исследовано по пять образцов из каждой партии сырья для определения разрывной нагрузки. Требуется выяснить, значимо ли влияют наладка машины и партии сырья на величину разрывной нагрузки.

По формуле (8.10) определяем средние значения. Табл. 8.11 преобразуется к следующему виду (табл. 8.12).

Таблица 8.12

Уровень наладки машины \ Партии сырья	B_1	B_2	$\bar{x}_{i\circ\circ}$
A_1	$\bar{x}_{11\circ} = 192$	$\bar{x}_{12\circ} = 170$	$\bar{x}_{1\circ\circ} = 181{,}0$
A_2	$\bar{x}_{21\circ} = 188$	$\bar{x}_{22\circ} = 202$	$\bar{x}_{2\circ\circ} = 195{,}0$
A_3	$\bar{x}_{31\circ} = 180$	$\bar{x}_{32\circ} = 164$	$\bar{x}_{3\circ\circ} = 172{,}0$
$\bar{x}_{\circ j\circ}$	$\bar{x}_{\circ 1\circ} = 186{,}7$	$\bar{x}_{\circ 2\circ} = 178{,}7$	182,7

Находим суммы квадратов (см. табл. 8.10): $Q_1 = 2686{,}6$; $Q_2 = 480$; $Q_3 = 1860$; $Q_4 = 22\,360$; $Q = Q_1 + Q_2 + Q_3 + Q_4 = 27386{,}7$.

Вычисляем оценки дисперсий: $s_1^2 = 2686{,}7$; $s_2^2 = 240$; $s_3^2 = 930$; $s^2 = 944{,}4$; $s_4^2 = 931{,}7$.

Табл. 8.10 в нашем случае преобразуется к следующему виду (табл. 8.13).

Таблица 8.13

Компонента дисперсии	Суммы квадратов	Число степеней свободы	Оценка дисперсии
Между средними по строкам (по фактору A)	2686,7	1	2686,7
Между средними по столбцам (по фактору B)	480	1	240
Взаимодействие	1 860	2	930
Остаточная	22 360	24	931,7
Полная	27386,7	29	944,4

Вычисляем отношения дисперсий: $F_A=2686{,}7/931{,}7=2{,}88$; $F_B=240/931{,}7=0{,}26$.

При уровне значимости $\alpha=0{,}05$; $k_4=24$ и $k_1=1$ степенями свободы для F_A и $k_2=24$, $k_2=2$ степеням свободы для F_B находим следующие табличные значения: $F_{\alpha(A)}=4{,}26$; $F_{\alpha(B)}=3{,}40$. Сравнивая табличные значения с вычисленными, имеем $F_A<F_{\alpha(A)}$; $F_B<F_{\alpha(B)}$. Следовательно, нулевая гипотеза о равенстве средних не отвергается, т. е. влияние фактора A (уровня наладки машины) и фактора B (партии сырья) на величину разрывной нагрузки незначимо.

§ 8.4. Дисперсионный анализ с неравным числом наблюдений в ячейке

В предыдущих параграфах настоящей главы рассматривались схемы дисперсионного анализа, когда в ячейках матрицы наблюдений находилось одинаковое число наблюдений. В рассмотренных случаях проверка гипотез о действии исследуемых факторов на результирующий признак сводилась к простым вычислительным схемам.

Однако в матрице наблюдений может оказаться неодинаковое количество наблюдений в ячейках, а некоторые ячейки вообще могут быть пустыми. Предположим, например, что на текстильной фабрике необходимо опре-

делить влияние качества сырья в партиях (фактор A) и состояния машин (фактор B) на механические свойства пряжи, а именно на удлинение (результирующий признак x). Взято три партии пряжи (три уровня фактора A) и пять машин (пять уровней фактора B), причем в каждую ячейку решено записать по пять наблюдений. Планируемая матрица наблюдений должна иметь вид табл. 8.14.

Таблица 8.14

A \ B	1	2	3	4	5
I	x_{111}, x_{112}, x_{113}, x_{114}, x_{115} (1,1)	x_{121}, x_{122}, x_{123}, x_{124}, x_{125} (1,2)	x_{131}, x_{132}, x_{133}, x_{134}, x_{135} (1,3)	x_{141}, x_{142}, x_{143}, x_{144}, x_{145} (1,4)	x_{151}, x_{152}, x_{153}, x_{154}, x_{155} (1,5)
II	x_{211}, x_{212}, x_{213}, x_{214}, x_{215} (2,1)	x_{221}, x_{222}, x_{223}, x_{224}, x_{225} (2,2)	x_{231}, x_{232}, x_{233}, x_{234}, x_{235} (2,3)	x_{241}, x_{242}, x_{243}, x_{244}, x_{245} (2,4)	x_{251}, x_{252}, x_{253}, x_{254}, x_{255} (2,5)
III	x_{311}, x_{312}, x_{313}, x_{314}, x_{315} (3,1)	x_{321}, x_{322}, x_{323}, x_{324}, x_{325} (3,2)	x_{331}, x_{332}, x_{333}, x_{334}, x_{335} (3,3)	x_{341}, x_{342}, x_{343}, x_{344}, x_{345} (3,4)	x_{351}, x_{352}, x_{353}, x_{354}, x_{355} (3,5)

Однако оказалось, что вторая машина испортилась при обработке первой партии сырья, а третья машина на некоторое время была остановлена для профилактического осмотра. В результате ячейка с индексом (1,2) осталась пустой, а для заполнения ячейки (1,3) имеются лишь три наблюдения из пяти запланированных, поскольку первая партия сырья была израсходована и получить недостающую информацию оказалось невозможным. Аналогичные ситуации возникли при обработке второй и третьей партий сырья, поэтому в результате эксперимента получили матрицу наблюдений (табл. 8.15).

Для того чтобы при обработке матрицы наблюдений можно было воспользоваться уже известными схемами, необходимо вычеркнуть столбцы, в которых встречаются пустые ячейки и ячейки с числом наблюдений, меньшим пяти. В этом случае получается матрица, не позволяю-

Таблица 8.15

A \ B	1	2	3	4	5
I	x_{111}, x_{112}, x_{113}, x_{114}, x_{115} (1, 1)	0 (1, 2)	x_{131}, x_{132}, x_{133} (1, 3)	x_{141}, x_{142}, x_{143}, x_{144}, x_{145} (1, 4)	x_{151}, x_{152}, x_{153}, x_{154}, x_{155} (1, 5)
II	x_{211}, x_{212}, x_{213}, x_{214}, x_{215} (2, 1)	x_{221}, x_{222}, x_{223}, x_{224}, x_{225} (2, 2)	x_{231}, x_{232}, x_{233}, x_{234}, x_{235} (2, 3)	x_{241}, x_{242}, x_{243} (2, 4)	0 (2, 5)
III	x_{311}, x_{312}, x_{313}, x_{314}, x_{315} (3, 1)	x_{321}, x_{322}, x_{323}, x_{324}, x_{325} (3, 2)	x_{331}, x_{332}, x_{333} (3, 3)	x_{341}, x_{342}, (3, 4)	x_{351}, x_{352} (3, 5)

щая решить поставленную задачу (определить влияние фактора B на результирующий признак x (табл. 8.16).

Таблица 8.16

A \ B	1
I	x_{111}, x_{112}, x_{113}, x_{114}, x_{115}
II	x_{211}, x_{212}, x_{213}, x_{214}, x_{215}
III	x_{311}, x_{312}, x_{313}, x_{314}, x_{315}

Для случая, когда в ячейках матрицы наблюдений находится неодинаковое их количество, разработана особая методика, с подробным изложением которой можно ознакомиться в работе [10]. Остановимся на двухфакторном дисперсионном анализе при наличии пустых ячеек в матрице наблюдений. Представим каждое наблюдение x_{ijk} в виде

$$x_{ijk} = \mu + \gamma_i + g_j + v_{ij} + e_{ijk},$$

где $i=1, 2,..., I$ (I — количество строк); $j=1, 2,..., J$ (J — количество столбцов); $k=1, 2,..., K$ (K — количество на-

блюдений в ячейке); μ — средняя всего комплекса; γ_i — эффект, обусловленный действием фактора A; g_j — эффект, обусловленный действием фактора B; ν_{ij} — эффект, обусловленный взаимодействием i-го уровня фактора A с j-м уровнем фактора B; e_{ijk} — вариация результатов внутри отдельной ячейки, с помощью которой учитываются все неконтролируемые факторы.

На составляющие наблюдения x_{ijk} накладывают ограничения. Так, предполагают, что множество $\{e_{ijk}\}$ подчиняется нормальному закону распределения с математическим ожиданием, равным нулю, и дисперсией σ^2. Средние эффекты факторов A, B и их взаимодействие также предполагают равными нулю, т. е.

$$\frac{1}{I}\sum_{i=1}^{I}\gamma_i=\frac{1}{J}\sum_{j=1}^{J}g_j=\frac{1}{I}\sum_{i=1}^{I}\nu_{ij}=\frac{1}{J}\sum_{j=1}^{J}\nu_{ij}=0.$$

Для примера запишем наблюдения ячеек (2,4) и (3,5) в табл. 8.15 во вновь принятых обозначениях:

$$x_{241}=\mu+\gamma_2+g_4+\nu_{24}+e_{241},\quad x_{242}=\mu+\gamma_2+g_4+\nu_{24}+e_{242},$$
$$x_{243}=\mu+\gamma_2+g_4+\nu_{24}+e_{243},\quad x_{351}=\mu+\gamma_3+g_5+\nu_{35}+e_{351},$$
$$x_{352}=\mu+\gamma_3+g_5+\nu_{35}+e_{352}.$$

Если в матрице наблюдений есть пустые ячейки, то анализ начинают с проверки отсутствия эффекта взаимодействия. Обычно эту гипотезу обозначают H_{AB}. Если гипотеза H_{AB} отвергается, т. е. взаимодействие значимо, то считают доказанным влияние и фактора A, и фактора B на результирующий признак x. Если же гипотеза H_{AB} принимается, т. е. взаимодействие факторов незначимо, то переходят к проверке влияния на результирующий признак каждого из факторов A и B по схеме однофакторного дисперсионного анализа.

Перейдем к проверке гипотезы H_{AB}. Общая идея проверки гипотезы H_{AB} сводится к следующему:

1. Сначала вычисляют сумму квадратов отклонений каждого наблюдения ячейки от своего среднего. Полученные результаты по всем непустым ячейкам складывают. Полученная сумма (обозначим ее через Q_e) характеризует вариацию наблюдений, которая вызвана неконтролируемыми факторами. Действительно, $M(e_{ijk})=0$,

поэтому среднюю $\bar{x}_{ij\bullet}$ ячейки (ij) можно представить в виде $\bar{x}_{ij\bullet}=\mu+\gamma_i+g_j+\nu_{ij}$, следовательно,

$$x_{ijk}-\bar{x}_{ij\bullet}=(\mu+\gamma_i+g_j+\nu_{ij}+e_{ijk})-$$
$$-(\mu+\gamma_i+g_j+\nu_{ij})=e_{ijk}.$$

Итак,

$$Q_e=\sum_{(ij)\in D}\sum_{k=1}^{K}(x_{ijk}-\bar{x}_{ij\bullet})^2,$$

где D — множество непустых ячеек. Стастистика Q_e имеет $k=n-p$ степеней свободы, где n — общее количество наблюдений; p — количество непустых ячеек. Рассчитать полученную статистику по матрице наблюдений не составляет труда.

2. Если взаимодействие факторов незначимо, то среднюю $(\bar{x}_{ij\bullet})$ ячейки можно представить в следующем виде: $\bar{x}_{ij\bullet}=\mu+\gamma_i+g_j$. Тогда

$$x_{ijk}-\bar{x}_{ij\bullet}=(\mu+\gamma_i+g_j+\nu_{ij}+e_{ijk})-$$
$$-(\mu+\gamma_i+g_j)=(\nu_{ij}+e_{ijk}).$$

Сумма квадратов отклонений каждого наблюдения ячейки от своего среднего

$$Q_{AB+e}=\sum_{ij\in D}\sum_{k=1}^{K}(\nu_{ij}+e_{ijk})^2=\sum_{ij\in D}\sum_{k=1}^{K}(x_{ijk}-\mu-\gamma_i-g_j)^2$$

содержит сумму эффектов взаимодействия и вариацию наблюдений, вызванную действием неконтролируемых факторов. Если из Q_{AB+e} вычесть Q_e, то получим эффект, характеризующий взаимодействие исследуемых факторов. Эта разность имеет $p-I-J+1$ степней свободы.

3. Если гипотеза H_{AB} справедлива, т. е. взаимодействие факторов незначимо, то отношение

$$\frac{(Q_{AB}+Q_e)/(p-I-J+1)}{Q_e/(n-p)}$$

имеет F-распределение с $p-I-J+1=k_1$ и $n-p=k_2$ степенями свободы.

Затем проверка гипотезы сводится к следующему: выбирают уровень значимости α, соответствующий про-

веряемой гипотезе H_{AB}. При данном уровне значимости α и числе степеней свободы k_1 и k_2 находят табличное значение F. Если $F_{табл} \geqslant F$, то нет оснований отвергать проверяемую гипотезу H_{AB}. Если $F_{табл} < F$, то гипотезу H_{AB} отвергают, т. е. взаимодействие факторов значимо.

4. Для того чтобы вычислить Q_{AB+e}, необходимо найти оценки μ, γ_i, g_j при гипотезе H_{AB}.

Известно, что наилучшие оценки для μ, γ_i, g_j можно получить по методу наименьших квадратов, т. е. минимизируя статистику $\sum\limits_{ij \in D} \sum\limits_{k=1}^{K_{ij}} (x_{ijk} - \mu - \gamma_i - g_j)^2$ по каждому из параметров μ, γ_i и g_j, а именно приравнивая нулю производные по каждому из параметров:

$$\frac{\partial Q_{AB+e}}{\partial \mu} = 0, \quad \frac{\partial Q_{AB+e}}{\partial \gamma_i} = 0, \quad \frac{\partial Q_{AB+e}}{\partial g_j} = 0,$$

$$i = 1, 2, \ldots, I, \quad j = 1, 2, \ldots, J. \tag{8.11}$$

Пусть G_i — количество элементов в i-м столбце. Тогда систему уравнений (8.11) можно записать в виде

$$\begin{cases} n\mu + \sum\limits_{i=1}^{I} \gamma G_i + \sum\limits_{j=1}^{J} g H_j = \sum\limits_{ij \in D} \sum\limits_{k=1}^{K} x_{ijk}, \\ G_i \mu + G_i \gamma_i + \sum\limits_{j=1}^{J} k_{ij} g_j = \sum\limits_{j=1}^{J} \sum\limits_{k=1}^{K} x_{ijk}, \\ H_j \mu + \sum\limits_{i=1}^{I} k_{ij} \gamma_i + H_j g_j = \sum\limits_{i=1}^{I} \sum\limits_{k=1}^{K} x_{ijk}. \end{cases} \tag{8.12}$$

Система (8.12) содержит $I+J+1$ уравнение с $I+J+1$ неизвестными (μ, γ_1, $\gamma_2, \ldots, \gamma_I$, $g_1, \ldots, g_J$). Однако среди уравнений этой системы лишь $I+J-1$ линейно независимы, так как сумма уравнений, отмеченных$_*$, равна первому уравнению, а сумма уравнений, отмеченных$_{**}$, также равна первому уравнению. Из этого следует, что решение системы не является единственным. Учитывая, что $(1/I)\sum\limits_{i=1}^{I} \gamma_i = 0$ и $(1/J)\sum\limits_{j=1}^{J} g_j = 0$, и исключая из системы (8.12) любые два уравнения (например, первое и последнее), получаем

$$\begin{cases} G_i\mu + G_i\gamma_i + \sum_{j=1}^{J} k_{ij} g_j = \sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}, \\ H_j\mu + \sum_{i=1}^{I} k_{ij}\gamma_i + H_j g_j = \sum_{i=1}^{I}\sum_{k=1}^{K} x_{ijk}, \\ \sum_{i=1}^{I}\gamma_i = 0, \ \sum_{j=1}^{J} g_j = 0. \end{cases} \qquad (8.13)$$

Система (8.13) имеет единственное решение. Решая систему, определяем μ, γ_i, g_j, а затем рассчитываем Q_{AB+e}.

В заключение остановимся на *критерии допустимого количества пустых ячеек*. Количество пустых ячеек считается допустимым, если числа степеней свободы $k_1 = p - I - J + 1$ и $k_2 = n - p$ отличны от нуля и при этом система (8.13) имеет единственное решение. При невыполнении хотя бы одного из этих условий задача оценки влияния действующих факторов не может быть решена.

Приведем пример. Рассмотрим двухфакторный комплекс. Матрица наблюдений имеет следующий вид:

A \ B	B_1	B_2	B_3
A_1	0,5; 0,6		0,4; 0,3
A_2	0,6; 0,6	0,2; 0,6; 0,4	0,6; 0,5

Для удобства вычислений составим вспомогательную таблицу, в каждой ячейке которой запишем сумму, среднее и количество элементов ячейки (ij).

Система (8.12) в данном случае принимает вид

$$\begin{cases} 11\mu + 4\gamma_1 + 7\gamma_2 + 4g_1 + 3g_2 + 4g_3 = 5{,}3, \\ 4\mu + 4\gamma_1 - 2g_1 + 2g_3 = 1{,}8, \\ 7\mu + 7\gamma_2 + 2g_1 + 3g_2 + 2g_3 = 3{,}5, \\ 4\mu + 2\gamma_1 + 2\gamma_2 + 4g_1 = 2{,}3, \\ 3\mu + 3\gamma_2 + 3g_2 = 1{,}2, \\ 4\mu + 2\gamma_1 + 2\gamma_2 + 4g_3 = 1{,}8. \end{cases}$$

Таблица 8.17

A \ B	B_1	B_2	B_3	Сумма строки	Средняя строки	Количество элементов в строке
A_1	1,1 0,55 2		0,7 0,35 2	1,8	0,45	4
A_2	1,2 0,6 2	1,2 0,4 3	1,1 0,55 2	3,5	0,5	7
Сумма столбца	2,3	1,2	1,8	5,3		
Средняя столбца	0,575	0,4	0,45		0,482	
Количество элементов в столбце	4	3	4			11

Заметим, что сумма второго и третьего уравнений есть первое уравнение; сумма четвертого, пятого и шестого уравнений также есть первое уравнение системы (8.13). В рассматриваемом примере система (8.13) преобразуется к виду

$$\begin{aligned}
4\mu + 4\gamma_1 \qquad\quad + 2g_1 \qquad\quad + 2g_3 &= 1{,}8,\\
7\mu + \qquad 7\gamma_2 + 2g_1 + 3g_2 + 2g_3 &= 3{,}5,\\
4\mu + 2\gamma_1 + 2\gamma_2 + 4g_1 \qquad\quad &= 2{,}3,\\
3\mu \qquad + 3\gamma_2 \qquad\quad + 3g_2 \qquad\quad &= 1{,}2,\\
\gamma_1 + \gamma_2 \qquad\quad \qquad\quad &= 0,\\
g_1 + g_2 + g_3 &= 0.
\end{aligned}$$

Для решения подобных систем можно использовать, например, метод обратных матриц. Однако в данном случае мы не воспользуемся общим методом. Преобразуем сначала второе и четвертое уравнения системы, для чего выпишем их отдельно и представим в следующем виде:

$$7\mu + 7\gamma_2 + 2\underbrace{(g_1+g_2+g_3)}_{0} + g_2 = 3{,}5,$$

$$\mu + \gamma_2 \qquad\quad + g_2 = 0{,}4.$$

Решим эти уравнения относительно g_2 и $\mu+\gamma_2$. В результате получим: $g_2=-0{,}117$, $\mu+\gamma_2=0{,}517$. Представим теперь первое уравнение в следующем виде:

$$4\mu + 4\gamma_1 + (2g_1 + 2g_2 + 2g_3) - 2g_2 = 1{,}8,$$

откуда $4\mu+4\gamma_1-2g_2=1{,}8$.

Используя вычисленные ранее значения g_2 и $\mu+\gamma_2$, находим: $\mu=0{,}454$, $\gamma_2=0{,}063$, $\gamma_1=-0{,}063$.

Из третьего уравнения находим $g_1=0{,}121$, из шестого уравнения определяем $g_3=0{,}004$.

Рассчитаем Q_{AB+e}. Заметим, что формулу для определения Q_{AB+e} можно преобразовать, возведя в квадрат выражение, стоящее в скобках, что позволит сократить количество вычислений. Имеем:

$$Q_{AB+e} = \sum_i \sum_j \sum_k x_{ijk}^2 - \mu \sum_i \sum_j \sum_k x_{ijk} - \sum_i \gamma_i \sum_j \sum_k x_{ijk} - \sum g_i \sum_j \sum_k \sum x_{ijk};$$

$$\sum_i \sum_j \sum_k x_{ijk}^2 = 2{,}75; \quad \sum_i \sum_j \sum_k x_{ijk} = 5{,}3;$$

$$\mu \sum_i \sum_j \sum_k x_{ijk} = 2{,}4062; \quad \sum_i \gamma_i \sum_j \sum_k x_{ijk} = 0{,}1069;$$

$$\sum_j g_j \sum_i \sum_k x_{ijk} = 0{,}1307; \quad Q_{AB+e} \approx 0{,}106.$$

Далее находим: $Q_e=0{,}095$; $F\approx 0{,}71$; $k_1=1$; $k_2=6$.

Пусть гипотеза H_{AB} проверяется при уровне значимости $\alpha=0{,}05$, тогда $F_{\text{табл}}=5{,}99$. Итак, $F_{\text{табл}}>F$, следовательно, взаимодействие исследуемых факторов незначимо, поэтому надо перейти к анализу каждого из факторов в отдельности по схеме дисперсионного анализа.

Глава 9

Корреляционный анализ

§ 9.1. О связях функциональных и статистических

Связи между различными явлениями в природе сложны и многообразны, однако их можно определенным образом классифицировать. В технике и естествознании часто речь идет о функциональной зависимости между переменными x и y, когда каждому возможному значению x поставлено в однозначное соответствие определенное значение y. Это может быть, например, зависимость между давлением и объемом газа (закон Бойля—Мариотта).

В реальном мире многие явления природы происходят в обстановке действия многочисленных факторов, влияние каждого из которых ничтожно, а число их велико. В этом случае связь теряет свою однозначность и изучаемая физическая система переходит не в определенное состояние, а в одно из возможных для нее состояний. Здесь речь может идти лишь о так называемой статистической связи. Статистическая связь состоит в том, что одна случайная переменная реагирует на изменение другой изменением своего закона распределения. Следовательно, для изучения статистической зависимости нужно знать аналитический вид двумерного распределения. Однако нахождение аналитического вида двумерного распределения по выборке ограниченного объема, во-первых, громоздко, во-вторых, может привести к значительным ошибкам. Поэтому на практике при исследовании зависимостей между случайными переменными X и Y обычно ограничиваются изучением зависимости между одной из них и условным математическим ожиданием другой, т. е. $M(Y|X=x)=\varphi(x)$ или $M(X|Y=y)=\varphi(y)$,

где $M(Y|X=x)$ — математическое ожидание случайной величины Y при условии, что случайная величина X приняла значение x. Аналогично для $M(X|Y=y)$.

Вопрос о том, что принять за зависимую переменную, а что — за независимую, следует решать применительно к каждому конкретному случаю.

Знание статистической зависимости между случайными переменными имеет большое практическое значение: с ее помощью можно прогнозировать значение зависимой случайной переменной в предположении, что независимая переменная примет определенное значение. Однако, поскольку понятие статистической зависимости относится к осредненным условиям, прогнозы не могут быть безошибочными. Применяя некоторые вероятностные методы, как будет показано далее, можно вычислить вероятность того, что ошибка прогноза не выйдет за определенные границы.

§ 9.2. Определение формы связи. Понятие регрессии

Определить форму связи — значит выявить механизм получения зависимой случайной переменной. При изучении статистических зависимостей форму связи можно характеризовать *функцией регрессии* (линейной, квадратной, показательной и т. д.).

Условное математическое ожидание $M(Y|X=x)$ случайной переменной Y, рассматриваемое как функция x, т. е. $M(Y|X=x)=\varphi(x)$, называется функцией регрессии случайной переменной Y относительно X (или функцией регрессии Y по X). Точно так же условное математическое ожидание $(M(X|\hat{Y}=y)$ случайной переменной X, т. е. $M(X|Y=y)=\varphi(y)$, называется функцией регрессии случайной переменной X относительно Y (или функцией регрессии X по Y).

Используя формулы § 3.9, на примере дискретного распределения найдем функцию регрессии.

Пример 9.1. Задана двумерная случайная переменная с плотностью вероятности

$$f(x,\ y)=\begin{cases}x+2y \text{ при } 0<x<1,\ 0<y<1,\\ 0 \text{ для всех остальных значений.}\end{cases}$$

Найти функцию регрессии случайной переменной Y для случая, когда X принимает значение x.

Решение. Для определения функции регрессии нужно знать условную плотность вероятности переменной Y при $X=x$. Имеем

$$f(Y \mid X=x)=\frac{f(x, y)}{\int\limits_0^1 f(x, y)\,\mathrm{d}y}=\frac{x+2y}{\int\limits_0^1 (x+2y)\mathrm{d}y}=\frac{x+2y}{x+1}.$$

Тогда функция регрессии

$$M(Y \mid X=x)=\int\limits_0^1 y f(Y \mid X=x)\,\mathrm{d}y=\int\limits_0^1 y\,\frac{x+2y}{x+1}\,\mathrm{d}y=\frac{0{,}5x+2/3}{x+1}.$$

Аналогично находится функция регрессии переменной X при $Y=y$.

Функция регрессии имеет важное значение при статистическом анализе зависимостей между переменными и может быть использована для прогнозирования одной из случайных переменных, если известно значение другой случайной переменной. Точность такого прогноза определяется дисперсией условного распределения.

Несмотря на важность понятия функции регрессии, возможности ее практического применения весьма ограничены. Как видно из приведенного примера, для оценки функции регрессии необходимо знать аналитический вид двумерного распределения (X, Y). Только в этом случае можно точно определить вид функции регрессии, а затем оценить параметры двумерного распределения. Однако для подобной оценки мы чаще всего располагаем лишь выборкой ограниченного объема, по которой нужно найти вид двумерного распределения (X, Y), а затем вид функции регрессии. Это может привести к значительным ошибкам, так как одну и ту же совокупность точек (x_i, y_i) на плоскости можно одинаково успешно описать с помощью различных функций. Именно поэтому возможности практического применения функции регрессии ограничены. Для характеристики формы связи при изучении зависимости используют понятие кривой регрессии.

Кривой регрессии Y по X (или Y на X) называют условное среднее значение случайной переменной Y, рассматриваемое как функция определенного класса, параметры которой находят методом наименьших квадратов по наблюденным значениям двумерной случайной величины (x, y), т. е.

$$\bar{y}(x)=\varphi(x, b_1, b_2, \ldots, b_m).$$

Аналогично определяется кривая регрессии X по Y (X на Y):

$$\bar{x}(y) = \varphi(y, c_1, c_2, \ldots, c_m).$$

Кривую регрессии называют также *эмпирическим уравнением* регрессии или просто *уравнением регрессии*. Уравнение регрессии является оценкой соответствующей функции регрессии.

Возникает вопрос: почему для определения кривой регрессии используют именно условное среднее $\bar{y}(x)$? Функция $\bar{y}(x)$ обладает одним замечательным свойством: она дает наименьшую среднюю погрешность оценки прогноза. Предположим, что кривая регрессии — произвольная функция. Средняя погрешность прогноза по кривой регрессии определяется математическим ожиданием квадрата разности между измеренной величиной и вычисленной по формуле кривой регрессии, т. е. $M[Y - \varphi(x)]^2$. Естественно потребовать вычисления такой кривой регрессии, средняя погрешность прогноза по которой была бы наименьшей. Таковой является $\varphi(x) = \bar{y}(x)$. Это следует из свойств минимальности рассеивания около центра распределения $\bar{y}(x)$. Если рассеивание вычисляется относительно $\varphi(x) \neq \bar{y}(x)$, то средний квадрат отклонения увеличивается. Поэтому можно сказать, что кривая регрессии, выражаемая как $\bar{y}(x)$, минимизирует среднюю квадратическую погрешность прогноза величины Y по X.

§ 9.3. Основные положения корреляционного анализа

Статистические связи между переменными можно изучать методом корреляционного и регрессионного анализа. С помощью этих методов решают разные задачи; требования, предъявляемые к исследуемым переменным, в каждом методе различны.

Основная задача корреляционного анализа — выявление связи между случайными переменными путем точечной и интервальной оценки парных коэффициентов корреляции, вычисления и проверки значимости множественных коэффициентов корреляции и детерминации, оценки частных коэффициентов корреляции. Корреляционный анализ позволяет также оценить функцию регрес-

сии одной случайной переменной на другую. Предпосылки корреляционного анализа следующие: 1) переменные величины должны быть случайными; 2) случайные величины должны иметь совместное нормальное распределение.

Рассмотрим простейший случай корреляционного анализа — двумерную модель. Введем основные понятия и опишем принцип проведения корреляционного анализа. Результаты этого параграфа легко обобщить на случай многомерной модели. Пусть X и Y — случайные переменные, имеющие совместное нормальное распределение. В этом случае, согласно § 3.9, связь между X и Y можно описать коэффициентом корреляции ρ. Этот коэффициент определяется как ковариация между X и Y, отнесенная к их средним квадратическим отклонениям:

$$\rho = \frac{K_{XY}}{\sigma_x \sigma_y}, \text{ или } \rho = M\left\{\frac{X - M(x)}{\sigma_x} \frac{Y - M(y)}{\sigma_y}\right\}. \qquad (9.1)$$

Оценкой коэффициента корреляции является *выборочный коэффициент корреляции r*. Для его нахождения необходимо знать оценки следующих параметров: $M(x)$, $M(y)$, σ_x, σ_y. Наилучшей оценкой математического ожидания является среднее ариметическое, т. е. $M(x) = \bar{x} = \sum_{i=1}^{n} x_i/n$. Оценкой дисперсии служит выборочная дисперсия, т. е.

$$\sigma_x^2 = s_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/n.$$

Тогда выборочный коэффициент корреляции

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y}. \qquad (9.2)$$

Коэффициент ρ называют также *парным коэффициентом корреляции*, а r — *выборочным парным коэффициентом корреляции*.

При совместном нормальном законе распределения случайных величин X и Y, используя рассмотренные выше параметры распределения и коэффициент корреляции, можно получить выражение для условного математического ожидания, т. е. записать выражение для функ-

ции регрессии одной случайной величины на другую. Так, функция регрессии Y на X имеет вид

$$M(Y \mid X = x) = M(y) + \rho \frac{\sigma_y}{\sigma_x} [X - M(x)]; \qquad (9.3)$$

функция регрессии X на Y — следующий вид:

$$M(X \mid Y = y) = M(x) + \rho \frac{\sigma_x}{\sigma_y} [Y - M(y)].$$

Выражения $\rho \frac{\sigma_y}{\sigma_x}$ и $\rho \frac{\sigma_x}{\sigma_y}$ называют *коэффициентами регрессии*. Подставив в (9.3) соответствующие оценки параметров, получим уравнения регрессии, график которых — прямая линия, проходящая через точку $C(\bar{x}; \bar{y})$.

Запишем уравнение регрессии y на x и x на y:

$$\bar{y}(x) = \bar{y} + r \frac{s_y}{s_x}(x - \bar{x}), \quad \bar{x}(y) = \bar{x} + r \frac{s_x}{s_y}(y - \bar{y}). \qquad (9.4)$$

Таким образом, в корреляционном анализе на основе оценок параметров двумерной нормальной совокупности получаем оценки тесноты связи между случайными переменными и можем оценить регрессию одной переменной на другую. Особенностью корреляционного анализа является строго линейная зависимость между переменными. Это обусловливается исходными предпосылками. На практике корреляционный анализ можно применять для обработки наблюдений, сделанных на предприятиях при нормальных условиях работы, если случайные изменения свойства сырья или других факторов вызывают случайные изменения свойств продукции.

§ 9.4. Свойства коэффициента корреляции

Коэффициент корреляции является одним из самых распространенных способов измерения связи между случайными переменными. Рассмотрим некоторые свойства этого коэффициента.

Теорема 9.1. *Коэффициент корреляции принимает значения на интервале* $(-1, +1)$.

Доказательство. Докажем справедливость утверждения для случая дискретных переменных. Запишем явно неотрицательное выражение:

$$\frac{1}{n} \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{\sigma_x} \pm \frac{y_i - \bar{y}}{\sigma_y} \right]^2 \geqslant 0.$$

Возведем выражение под знаком суммы в квадрат:

$$\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{\sigma_x^2}\pm 2\,\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y}+\right.$$

$$\left.+\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{\sigma_y^2}\right\}\geqslant 0.$$

Первое и третье из слагаемых равны единице, поскольку из определения дисперсии следует, что $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\sigma_x^2$ и $\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2=\sigma_y^2$. Таким образом, окончательно получаем $1\pm 2\rho+1\geqslant 0$, откуда $-1\leqslant\rho\leqslant+1$.

Если коэффициент корреляции положителен, то связь между переменными также положительна и значения переменных увеличиваются или уменьшаются одновременно. Если коэффициент корреляции имеет отрицательное значение, то при увеличении одной переменной уменьшается другая.

Приведем следующее важное свойство коэффициента корреляции: *коэффициент корреляции не зависит от выбора начала отсчета и единицы измерения, т. е. от любых постоянных* a_1 *и* a_2, b_1 *и* b_2 *таких, что* $a_1>0$ *и* $a_2>0$, *т. е.*

$$\rho(a_1X+b_1;\ a_2Y+b_1)=\rho_{xy}.$$

Таким образом, переменные X и Y можно уменьшать или увеличивать в a раз, а также вычитать или прибавлять к значениям X и Y одно и то же число b. В результате величина коэффициента корреляции не изменится.

Если коэффициент корреляции $\rho_{xy}=0$, то случайные переменные некоррелированы. Понятие некоррелированности не следует смешивать с понятием независимости, независимые величины всегда некоррелированы. Однако обратное утверждение невероятно: некоррелированные величины могут быть зависимы и даже функционально, однако эта связь не линейная.

Выборочный коэффициент корреляции вычисляют по формуле (9.2). Имеется несколько модификаций этой формулы, которые удобно использовать при той или иной форме представления исходной информации. Так,

при малом числе наблюдений выборочный коэффициент корреляции удобно вычислять по формуле

$$r=\frac{n\sum\limits_{x}xy-\sum\limits_{x}x\sum\limits_{y}y}{\sqrt{n\sum\limits_{x}x^2-\left(\sum\limits_{x}x\right)^2}\sqrt{n\sum\limits_{y}y^2-\left(\sum\limits_{y}y\right)^2}}.$$

Если информация имеет вид корреляционной таблицы (см. § 9.5), то удобно пользоваться формулой

$$r=\frac{\sum\limits_{x,y}xym_{xy}-\sum\limits_{x}xm_x\sum\limits_{y}ym_y/\sum\limits_{x,y}m_{xy}}{\sum\limits_{x,y}m_{xy}s_xs_y}, \qquad (9.5)$$

где $\Sigma m_x=\Sigma m_y=\Sigma m_{xy}=n$; m_x — суммарная частота наблюдаемого значения признака x при всех значениях y; m_y — суммарная частота наблюдаемого значения признака y при всех значениях x; m_{xy} — частота появления пары признаков (x, y).

Из формулы (9.2) очевидно, что $r_{xy}=r_{yx}$, т. е. величина выборочного коэффициента корреляции не зависит от порядка следования переменных, поэтому обычно пишут просто r.

§ 9.5. Поле корреляции. Вычисление оценок параметров двумерной модели

На практике для вычисления оценок параметров двумерной модели удобно использовать *корреляционную таблицу* и *поле корреляции*. Пусть, например, изучается зависимость между объемом выполненных работ (y) и накладными расходами (x). Имеем выборку из генеральной совокупности, состоящую из 150 пар переменных (x_i, y_i). Считаем, что предпосылки корреляционного анализа выполнены.

Пару случайных чисел (x_i, y_i) можно изобразить графически в виде точки с координатами $(x_i; y_i)$. Аналогично можно изобразить весь набор пар случайных чисел (всю выборку). Однако при большом объеме выборки это затруднительно. Задача упрощается, если выборку упорядочить, т. е. переменные сгруппировать. Сгруппированные ряды могут быть как дискретными, так и интервальными.

По осям координат откладывают или дискретные значения переменных, или интервалы их изменения. Для интервального ряда наносят координатную сетку. Каждую пару переменных из данной выборки изображают в виде точки с соответствующими координатами для дискретного ряда или в виде точки в соответствующей клетке для интервального ряда. Такое изображение корре-

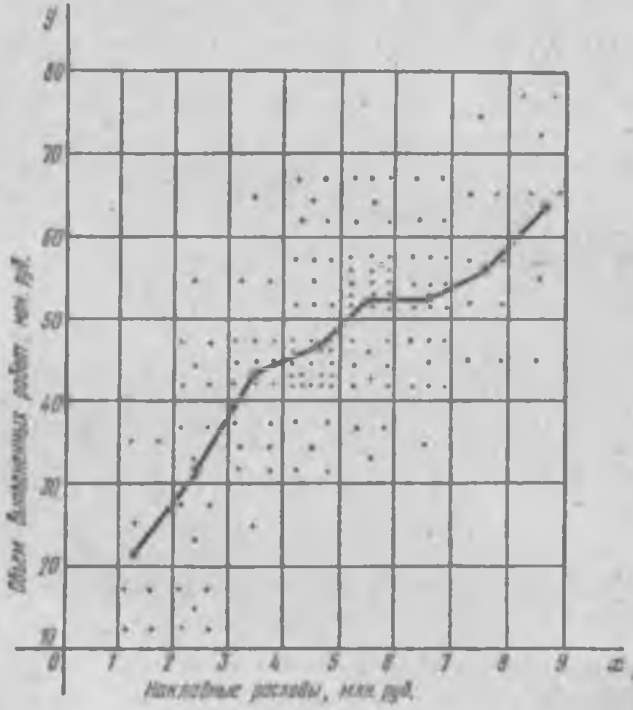

Рис. 9.1

ляционной зависимости называют *полем корреляции*. На рис. 9.1 изображено поле корреляции для выборки, состоящей из 150 пар переменных (ряд интервальный). Если вычислить средние значения y в каждом интервале изменения x [обозначим их $y_i(x)$], нанести эти точки на рис. 9.1 и соединить между собой, то получим ломаную линию, по виду которой можно судить, как в среднем меняются y в зависимости от изменения x. По виду этой линии можно также сделать предположение о форме связи между переменными. В данном случае ломаную линию можно аппроксимировать прямой линией, так как она достаточно хорошо приближается к ней.

Таблица 9.1

Объем выполненных работ y, млн. руб. \ Накладные расходы x, млн. руб.	1—2 1,5	2—3 2,5	3—4 3,5	4—5 4,5	5—6 5,5	6—7 6,5	7—8 7,5	8—9 8,5	m_y
10—20 15	4	5							9
20—30 25	1	3	1						5
30—40 35	2	3	6	5	3	1			20
40—50 45		5	9	19	8	7	2	1	51
50—60 55		1	2	7	16	9	4	2	41
60—70 65			1	5	6	4	2	2	20
70—80 75							1	3	4
m_x	7	17	19	36	33	21	9	8	150

По выборочным данным можно построить также корреляционную табл. 9.1. Корреляционную таблицу, как и поле корреляции, строят по сгруппированному ряду (дискретному или интервальному). Табл. 9.1 построена на основе интервального ряда. В первой строке и первом столбце таблицы помещают интервалы изменения x и y и значения середин интервалов. Так, например, 1,5 — середина интервала изменения $x=1\div2{,}15$ — середина интервала изменения $y=10\div20$. В ячейки, образованные пересечением строк и столбцов, заносят частоты попадания пар значений (x, y) в соответствующие интервалы по x и y. Например, частота 4 означает, что в интервал изменения y от 10 до 20 попало 4 пары наблюдавшихся значений. Эти частоты обозначают m_{xy}. В 9-й строке и 10-м столбце находятся значения m_x и m_y — суммы m_{xy} по соответствующим столбцу и строке.

Как будет показано в дальнейшем, корреляционной таблицей удобно пользоваться при вычислении коэффициентов корреляций и параметров уравнений регрессии.

Корреляционная таблица построена на основе интервального ряда, поэтому для оценок параметров воспользуемся формулами гл. I для вычисления средней арифметической и дисперсии. Имеем:

$$\bar{x} = \frac{\sum\limits_x x m_x}{\sum\limits_x m_x}; \quad \bar{y} = \frac{\sum\limits_y y m_y}{\sum\limits_y m_y}; \tag{9.6}$$

$$s_x^2 = \frac{\sum\limits_x x^2 m_x}{\sum\limits_x m_x} - \left(\frac{\sum\limits_x x m_x}{\sum\limits_x m_x}\right)^2; \quad s_y^2 = \frac{\sum\limits_y y^2 m_y}{\sum\limits_y m_y} - \left(\frac{\sum\limits_y y m_y}{\sum\limits_y m_y}\right)^2.$$

Пример 9.2. По данным, приведенным в табл. 9.1, вычислить оценки параметров двумерной модели; записать уравнения регрессии.

Решение. Вычислим все необходимые для расчета суммы по формулам (9.4)—(9.6) и запишем их в табл. 10.1.

Найдем оценки параметров $\bar{x}$, $\bar{y}$, s_x, s_y. Получаем:

$$\bar{x} = 735/150 = 4{,}9; \quad \bar{y} = 7110/150 = 47{,}4,$$

$$s_x = \sqrt{4053{,}5/150 - (735/150)^2} = 1{,}74,$$

$$s_y = \sqrt{363950/150 - (7110/150)^2} = 13{,}4.$$

Вычислим выборочный коэффициент корреляции:

$$r = (37175{,}0 + 735 \cdot 7110/150)/150 \cdot 1{,}74 \cdot 13{,}4 = 0{,}67.$$

Теперь запишем уравнение регрессии Y на X и X на Y:

$$\bar{y}(x) = 47{,}4 + 0{,}67 \cdot 13{,}4/1{,}74\,(x - 4{,}9) = 47{,}4 + 5{,}16\,(x - 4{,}9),$$

$$\bar{x}(y) = 4{,}9 + 0{,}67 \cdot 1{,}74/13{,}4\,(y - 47{,}4) = 4{,}9 + 0{,}09\,(y - 47{,}4).$$

Полученные уравнения регрессии показывают, как в среднем изменяется Y (или X) в зависимости от изменения аргумента X (или Y). Задавая значения X (или Y), можно рассчитать ожидаемое значение Y (или X).

§ 9.6. Проверка гипотезы о значимости коэффициента корреляции

На практике коэффициент корреляции ρ обычно неизвестен. По результатам выборки может быть найдена его точечная оценка — выборочный коэффициент корреляции r.

Равенство нулю выборочного коэффициента корреляции еще не свидетельствует о равенстве нулю самого коэффициента корреляции, а следовательно, о некоррелированности случайных величин X и Y. Чтобы выяснить, находятся ли случайные величины в корреляционной зависимости, нужно проверить значимость выборочного коэффициента корреляции r, т. е. установить, достаточна ли его величина для обоснованного вывода о наличии корреляционной связи. Для этого проверяют нулевую гипотезу $H_0: \rho=0$. Предполагается наличие двумерного нормального распределения случайных переменных; объем выборки может быть любым. Вычисляют статистику $t=r\sqrt{(n-2)/(1-r^2)}$, которая имеет распределение Стьюдента с $k=n-2$ степенями свободы. Для проверки нулевой гипотезы по уровню значимости α и числу степеней свободы k находят по таблицам распределения Стьюдента (t-распределение; см. табл. 5 приложения) критическое значение $t_{\alpha, k}$, удовлетворяющее условию $P(|t|\geqslant t_{\alpha, k})=\alpha$. Если $|t|\geqslant t_{\alpha, k}$, то нулевую гипотезу об отсутствии корреляционной связи между переменными X и Y следует отвергнуть. Переменные считают зависимыми. При $|t|<t_{\alpha, k}$ нет оснований отвергать нулевую гипотезу.

В случае значимого выборочного коэффициента корреляции есть смысл построить доверительный интервал для коэффициента корреляции ρ. Однако для этого нужно знать закон распределения выборочного коэффициента корреляции r.

Плотность вероятности выборочного коэффициента корреляции имеет сложный вид, поэтому прибегают к специально подобранным функциям от выборочного коэффициента корреляции, которые сводятся к хорошо изученным распределениям, например нормальному или Стьюдента. Чаще всего для подбора функции применяют преобразование Фишера. Вычисляют статистику:

$$z=\tfrac{1}{2}\ln\left(\frac{1+r}{1-r}\right),$$

где $r=\operatorname{th} z$ — гиперболический тангенс от z.

Распределение статистики z хорошо аппроксимируется нормальным распределением с параметрами

$$M(z)=\tfrac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)+\frac{\rho}{2(n-1)},\quad \sigma_z^2=\frac{1}{n-3}.$$

В этом случае доверительный интервал для ρ имеет вид $\text{th}\, z_1 < \rho < \text{th}\, z_2$. Величины $\text{th}\, z_1$ и $\text{th}\, z_2$ находят по таблицам по следующим значениям:

$$z_1 = {}^1/_2 \ln \frac{1+r}{1-r} - N_q \frac{1}{\sqrt{n-3}},$$

$$z_2 = {}^1/_2 \ln \frac{1+r}{1-r} + N_q \frac{1}{\sqrt{n-3}},$$

где N_q — нормированная функция Лапласа для q % доверительного интервала (см. табл. 2 приложений).

Если коэффициент корреляции значим, то коэффициенты регрессии также значимо отличаются от нуля, а интервальные оценки для них можно получить по следующим формулам:

$$b_{yx} - t_{\alpha,k} \frac{s_y \sqrt{1-r^2}}{s_x \sqrt{n-2}} < \beta_{YX} < b_{yx} + t_{\alpha,k} \frac{s_y \sqrt{1-r^2}}{s_x \sqrt{n-2}}, \quad (9.7)$$

$$b_{xy} - t_{\alpha,k} \frac{s_x \sqrt{1-r^2}}{s_y \sqrt{n-2}} < \beta_{XY} < b_{xy} + t_{\alpha,k} \frac{s_x \sqrt{1-r^2}}{s_y \sqrt{n-2}}, \quad (9.8)$$

где $t_{\alpha,k}$ имеет распределение Стьюдента с $k=n-2$ степенями свободы.

§ 9.7. Корреляционное отношение

На практике часто предпосылки корреляционного анализа нарушаются: один из признаков оказывается величиной не случайной, или признаки не имеют совместного нормального распределения. Однако статистическая зависимость между ними существует. Для изучения связи между признаками в этом случае существует общий показатель зависимости признаков, основанный на показателе изменчивости — общей (или полной) дисперсии.

Полной называется дисперсия признака относительно его математического ожидания. Так, для признака Y это $\sigma_Y^2 = M[Y - M(Y)]^2$. Дисперсию σ_Y^2 можно разложить на две составляющие, одна из которых характеризует влияние фактора X на Y, другая — влияние прочих факторов. Очевидно, чем меньше влияние прочих факторов, тем теснее связь, тем более приближается она к функциональной. Представим σ_Y^2 в следующем виде:

$$\sigma_Y^2 = M[M(Y \mid X=x) - M(Y)]^2 + M[Y - M(Y|X=x)]^2. \quad (9.9)$$

Первое слагаемое обозначим $\delta^2_{Y|X}$. Это дисперсия функции регрессии относительно математического ожидания признака (в данном случае признака Y); она измеряет влияние признака X на Y. Второе слагаемое обозначим $\sigma^2_{Y|X}$. Это дисперсия признака Y относительно функции регрессии. Ее называют также *средней* из условных дисперсий или *остаточной* дисперсией; $\sigma^2_{Y|X}$ измеряет влияние на Y прочих факторов.

Покажем, что σ_Y^2 действительно можно разложить на два таких слагаемых:

$$\sigma_Y^2 = M[Y - M(Y)]^2 = M[Y - M(Y \mid X = x) +$$
$$+ M(Y \mid X = x) - M(Y)]^2 = M[Y - M(Y \mid X = x)]^2 +$$
$$+ M[M(Y \mid X = x) - M(Y)]^2 + 2M\{[Y - M(Y \mid X = x)] \times$$
$$\times [M(Y \mid X = x) - M(Y)]\}. \quad (9.10)$$

Для простоты полагаем распределение дискретным. Имеем

$$M\{[Y - M(Y \mid X = x)][M(Y \mid X = x) - M(Y)]\} =$$
$$= \sum_x \sum_y [y - \bar{y}(x)][\bar{y}(x) - M(Y)] f(x, y) =$$
$$= \sum_x [\bar{y}(x) - M(Y)] f(x) \sum_y [Y - \bar{y}(x)] f(y \mid x) = 0,$$

так как при любом x справедливо равенство

$$\sum_y [y - \bar{y}(x)] f(y \mid x) = \sum_y y f(y \mid x) - \bar{y}(x) \sum_y f(y \mid x) =$$
$$= \bar{y}(x) - \bar{y}(x) = 0.$$

Третье слагаемое в равенстве (9.10) равно нулю, поэтому равенство (9.9) справедливо. Поскольку второе слагаемое в равенстве (9.9) оценивает влияние признака X на Y, то его можно использовать для оценки тесноты связи между X и Y. Тесноту связи удобно оценивать в единицах общей дисперсии σ_Y^2, т. е. рассматривать отношение $\{M[\bar{y}(x) - M(Y)]^2\}/\sigma_Y^2$. Эту величину обозначают $\eta^2_{TY|x}$ и называют *теоретическим корреляционным* отношением. Таким образом,

$$\eta^2_{TY|x} = \frac{M[\bar{y}(x) - M(Y)]^2}{\sigma^2_Y} = \frac{\delta^2_{Y|x}}{\sigma^2_Y}. \qquad (9.11)$$

Разделив обе части равенства (9.9) на σ^2_Y, получим

$$1 = (\sigma^2_{Y|x} + \delta^2_{Y|x})/\sigma^2_Y, \text{ или } 1 = \sigma^2_{Y|x}/\sigma^2_Y + \eta^2_{TY|x}.$$

Из последней формулы имеем

$$\eta^2_{TY|x} = 1 - \sigma^2_{Y|x}/\sigma^2_Y. \qquad (9.12)$$

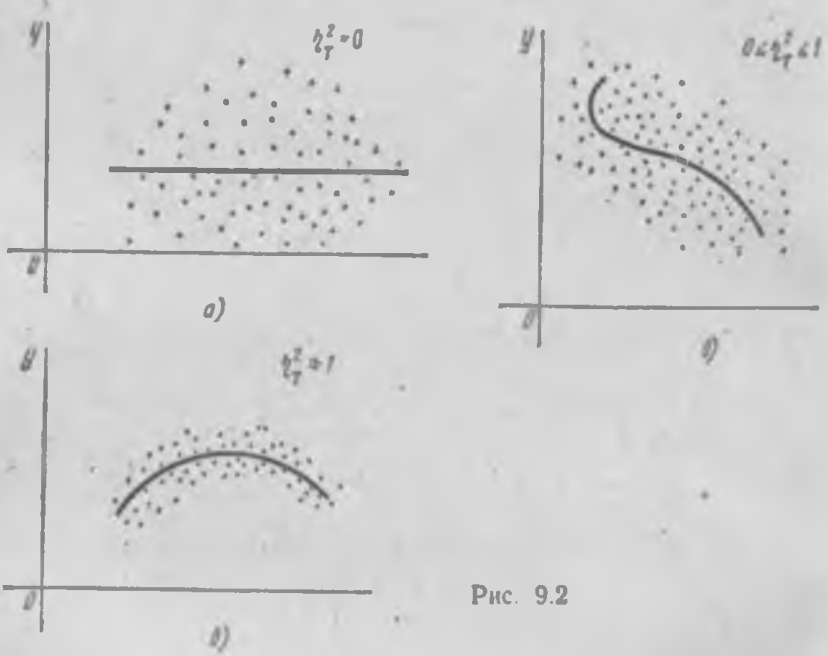

Рис. 9.2

Поскольку $\sigma^2_{Y|x} \leqslant \sigma^2_Y$, так как $\sigma^2_{Y|x}$ — составная часть σ^2_Y, то из равенства (9.12) следует, что значение $\eta^2_{TY|x}$ всегда заключено между нулем и единицей.

Все сделанные выводы справедливы и для $\eta^2_{Tx|y}$. Из равенства (9.12) следует, что $\eta^2_{TY|x} = 1$ только тогда, когда $\sigma^2_{Y|x} = 0$, т. е. отсутствует влияние прочих факторов и все распределение сконцентрировано на кривой регрессии $\bar{y}(x)$. В этом случае между Y и X существует функциональная зависимость. Далее, из равенства (9.12) следует, что $\eta^2_{TY|x} = 0$ тогда и только тогда, когда $\bar{y}(x) = M(Y) = \text{const}$, т. е. линия регрессии Y по X — горизонтальная прямая, проходящая через центр распреде-

ления. В этом случае можно сказать, что переменная Y не коррелирована с X (рис. 9.2, *а*, *б*, *в*). Аналогичными свойствами обладает $\eta^2_{TX|y}$ — показатель тесноты связи между X и Y.

Часто используют величину

$$\eta_{TY|x} = \sqrt{\eta^2_{TY|x}} = \delta_{y|x}/\sigma_Y. \qquad (9.13)$$

Считают, что она не может быть отрицательной. Значения величины $\eta_{TY|x}$ (или $\eta_{TX|y}$) также могут находиться лишь в пределах от нуля до единицы. Это очевидно из формулы (9.13).

Значения η^2, лежащие в интервале $0<\eta^2<1$, являются показателями тесноты группировки точек около кривой регрессии независимо от ее вида (формы связи). Корреляционное отношение η^2 связано с ρ^2 следующим образом: $0\leqslant\rho^2\leqslant\eta^2\leqslant 1$. В случае линейной зависимости между переменными $\rho^2=\eta^2$. Разность $\eta^2-\rho^2$ может быть использована как показатель нелинейности связи между переменными.

При вычислении η^2 по выборочным данным получаем *выборочное корреляционное отношение*. Обозначим его $\hat{\eta}^2$. Вместо дисперсий в этом случае используются их оценки. Тогда формула (9.12) принимает вид

$$\hat{\eta}^2 = s^2_{Y|x}/s^2_y.$$

Пример 9.3. По данным, приведенным в табл. 10.3, вычислить выборочное корреляционное отношение.

Решение. Вычислим $\hat{\eta}^2$ по формуле, приведенной выше. Представим s^2_y и $s^2_{Y|x}$ в следующем виде:

$$s^2_y = \frac{\sum\limits_y y^2 m_y}{\sum\limits_y m_y} - \left(\frac{\sum\limits_y y m_y}{\sum\limits_y m_y}\right)^2,$$

$$s^2_{Y|x} = \frac{1}{\sum\limits_{x,y} m}\,\Sigma\,[\bar{y}\,(x)-\bar{y}]^2\,m_x.$$

Суммы, необходимые для вычислений, приведены в табл. 10.3 и 9.2. Имеем $s^2_y=0{,}6746$.

Найдем $\bar{y}$:

$$\bar{y} = \sum_y y m_y / \sum_y m_y = 4{,}824.$$

Имеем $\overline{s^2_{Y\,x}} = 0{,}5528$. Таким образом, $\eta^2 = 0{,}8194$.

Таблица 9.2

$\bar{y}(x)$	$\bar{y}(x) - \bar{y}$	m_x	$[\bar{y}(x) - \bar{y}]^2 m_x$
6,06	1,23	8	12,1032
5,29	0,47	9	1,9881
5,20	0,38	6	0,8664
4,50	—0,32	8	0,8192
4,53	—0,29	6	0,5046
4,16	—0,66	5	2,1780
3,80	—1,02	4	4,1616
3,70	—1,12	4	5,0176
Σ		50	27,6387

§ 9.8. Понятие о многомерном корреляционном анализе

Частный коэффициент корреляции. Основные понятия корреляционного анализа, введенные для двумерной модели, можно распространить на многомерный случай. Задачи и предпосылки корреляционного анализа были сформулированы в § 9.3. Однако если при изучении взаимосвязи переменных по двумерной модели мы ограничивались рассмотрением парных коэффициентов корреляции, то для многомерной модели этого недостаточно. Многообразие связей между переменными находит отражение в *частных* и *множественных коэффициентах корреляции.*

Пусть имеется многомерная нормальная совокупность с m признаками X_1, X_2, ..., X_j, ..., X_m. В этом случае взаимозависимость между признаками можно описать *корреляционной матрицей.* Под корреляционной матрицей будем понимать матрицу, составленную из парных коэффициентов корреляции (вычисляются по формуле (9.1)):

$$Q_m = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1m} \\ \rho_{21} & 1 & \cdots & \rho_{2m} \\ \cdots & \cdots & \rho_{jk} & \cdots \\ \rho_{m1} & \rho_{m2} & \cdots & 1 \end{bmatrix} \qquad (9.14)$$

где ρ_{jk} — парные коэффициенты корреляции; m — порядок матрицы.

Оценкой парного коэффициента корреляции является выборочный парный коэффициент корреляции, определяемый по формуле (9.2), однако для m признаков формула (9.2) принимает вид

$$r_{jk} = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{n s_{x_j} s_{x_k}} \quad (j, k = 1,2, \ldots, m), \quad (9.15)$$

где j, k — порядковые номера признаков.

Как и в двумерном случае, для оценки коэффициента корреляции необходимо оценить математические ожидания и дисперсии. В многомерном корреляционном анализе имеем m математических ожиданий и m дисперсий, а также $m(m-1)/2$ парных коэффициентов корреляции. Таким образом, нужно произвести оценку $2m+$, $+m(m-1)/2$ параметров.

В случае многомерной корреляции зависимости между признаками более многобразны и сложны, чем в двумерном случае. Одной корреляционной матрицей нельзя полностью описать зависимости между признаками. Введем понятие частного коэффициента корреляции l-го порядка.

Пусть исходная совокупность состоит из m признаков. Можно изучать зависимости между двумя из них при фиксированном значении l признаков из $m-2$ оставшихся. Рассмотрим, например, систему из 5 признаков. Изучим зависимости между X_1 и X_2 при фиксированном значении признака X_3. В этом случае имеем частный коэффициент корреляции первого порядка, так как фиксируем только один признак.

Рассмотрим более подробно структуру частных коэффициентов корреляции на примере системы из трех признаков X_1, X_2, X_3. Эта система позволяет изучить частные коэффициенты корреляции только первого порядка, так как нельзя фиксировать больше одного признака. Частный коэффициент корреляции первого порядка для признаков X_1 и X_2 при фиксированном значении X_3 выражается через парные коэффициенты корреляции и имеет вид

$$\rho_{12\,3} = (\rho_{12} - \rho_{13}\rho_{23}) / \sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}\ . \quad (9.16)$$

Частный коэффициент корреляции, так же как и парный коэффициент корреляции, изменяется от -1 до $+1$. В общем виде, когда система состоит из m признаков, частный коэффициент корреляции l-го порядка может быть найден из корреляционной матрицы. Если $l=m-2$, то рассматривается матрица порядка m, при $l<m-2$ — подматрица порядка $l+2$, составленная из элементов матрицы Q_m, которые отвечают индексам коэффициента частной корреляции. Например, корреляционная матрица системы из пяти признаков имеет вид

$$Q_5 = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} & \rho_{25} \\ \rho_{32} & \rho_{32} & 1 & \rho_{34} & \rho_{35} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 & \rho_{45} \\ \rho_{51} & \rho_{52} & \rho_{53} & \rho_{54} & 1 \end{bmatrix}.$$

Для определения частного коэффициента корреляции второго порядка, например $\rho_{12.45}$, следует использовать подматрицу четвертого порядка, вычеркнув из исходной матрицы Q_5 третью строку и третий столбец, так как признак X_3 не рассматривают.

В общем виде формулу частного коэффициента корреляции l-го порядка ($l=m-2$) можно записать в виде

$$\rho_{jk\cdot 1,2,\ldots,m} = Q_{jk}/(Q_{jj}Q_{kk})^{1/2}, \qquad (9.17)$$

где Q_{jk} — алгебраические дополнения к элементу ρ_{jk} корреляционной матрицы Q_m; Q_{jj} и Q_{kk} — алгебраические дополнения к элементам ρ_{jj} и ρ_{kk} корреляционной матрицы Q_m.

Очевидно, что выражение (9.16) является частным случаем выражения (9.17), в чем легко убедиться, рассмотрев корреляционную матрицу Q_3.

Оценкой частного коэффициента корреляции l-го порядка является *выборочный частный коэффициент корреляции l-го порядка*. Он вычисляется на основе корреляционной матрицы, составленной из выборочных парных коэффициентов корреляции:

$$q_m = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{21} & 1 & \cdots & r_{2m} \\ & & r_{jk} & \\ r_{m1} & r_{m2} & \cdots & 1 \end{bmatrix}. \qquad (9.18)$$

Формула выборочного частного коэффициента корреляции имеет вид

$$r_{jk\,1,2,\ldots,m} = q_{jk}/(q_{jj}\,q_{kk})^{1/2}, \qquad (9.19)$$

где q_{jk}, q_{jj}, q_{kk} — алгебраические дополнения к соответствующим элементам матрицы (9.18).

Частный коэффициент корреляции l-го порядка, вычисленный на основе n наблюдений над признаками, имеет такое же распределение, что и парный коэффициент корреляции, вычисленный по $n-l$ наблюдениям. Поэтому значимость частных коэффициентов корреляции оценивают так же, как и в § 9.6.

Пример 9.4. Дана матрица выборочных парных коэффициентов корреляции размерности $m=4$ $(n=50)$:

$$R_4 = \begin{bmatrix} 1 & -0{,}85 & -0{,}62 & -0{,}21 \\ -0{,}85 & 1 & -0{,}53 & 0{,}34 \\ -0{,}62 & 0{,}53 & 1 & 0{,}46 \\ -0{,}21 & 0{,}34 & 0{,}46 & 1 \end{bmatrix}.$$

Вычислить оценки частных коэффициентов корреляции первого и второго порядков.

Решение. Для вычисления $r_{12.3}$ воспользуемся формулой (9.16), заменив значения параметров их оценками:

$$r_{12.3} = -0{,}85 - (-0{,}62 \cdot 0{,}53)/\sqrt{[1-(-0{,}62)^2]\,[1-0{,}53^2]} = \\ = -0{,}78.$$

Оценим значимость вычисленного коэффициента корреляции. Проверим нулевую гипотезу $H_0: \rho_{12.3}=0$. Уровень значимости $\alpha = 0{,}05$, число степеней свободы $k=(n-l)-2=47$. Вычислим статистику: $t = r\sqrt{(n-2)/(1-r^2)} = -8{,}544$. По табл. 5 приложений находим критическое значение $t_{0{,}05;\,47}=2{,}00$; имеем $|-8{,}544| > 2{,}00$. Гипотезу о равенстве нулю коэффициента корреляции следует отвергнуть. Можно говорить о наличии корреляционной связи между X_1 и X_2 при фиксированном значении X_3.

Вычислим $r_{23.1}$:

$$r_{23.1} = 0{,}53 - (-0{,}85 - 0{,}62)/\sqrt{[1-(20{,}85)^2]\,[1-0{,}62^2]} = 0{,}007.$$

Оценим значимость коэффициента корреляции. Имеем: $H_0 = \rho_{23.1}=0$, $k=(n-l)-2=47$, $\alpha=0{,}05$; статистика $t=0{,}0048$; критическое значение $t_{0{,}05;\,47}=2{,}00$; $|0{,}0048| < 2{,}00$. Нет оснований отвергать нулевую гипотезу. Величина $r_{23.1}$ показывает, что при фиксированном значении признака X_1 корреляция между X_2 и X_3 отсутствует.

Для нахождения $r_{12.34}$ воспользуемся формулой (9.19), которая в нашем случае имеет вид

$$r_{12.34} = q_{12}/\sqrt{q_{11}\,q_{22}}$$

Вычислим алгебраические дополнения к элементам матрицы r_{12}, r_{11}, r_{22}. Для элемента r_{12} алгебраическим дополнением является определитель

$$q_{12} = (-1)^3 \begin{vmatrix} -0,85 & 0,53 & 0,34 \\ -0,62 & 1 & 0,46 \\ -0,21 & 0,46 & 1 \end{vmatrix}.$$

Далее находим $q_{12}=0,42$. Аналогично, $q_{11}=0,56$, $q_{22}=0,48$, откуда $r_{12.34}=0,81$.

Оценим значимость вычисленного коэффициента корреляции: $H_0: \rho_{12.34}=0$, $k=(n-l)-2=46$, $\alpha=0,05$, $t=9,37$; критическое значение $t_{0,05;\,46}=2,00$, $|9,37|>2,00$. Нулевую гипотезу отвергаем. Можно говорить о наличии корреляционной связи между признаками X_1 и X_2 при фиксированном значении признаков X_3 и X_4.

Множественный коэффициент корреляции. Часто представляет интерес оценить связь одного из признаков со всеми остальными. Это можно сделать с помощью множественного, или *совокупного,* коэффициента корреляции

$$\rho_{j\cdot 1,2,\ldots,m} = \sqrt{1-|Q_m|/Q_{jj}}, \qquad (9.20)$$

где $|Q_m|$ — определитель корреляционной матрицы Q_m; Q_{jj} — алгебраическое дополнение к элементу ρ_{jj}.

Квадрат коэффициента множественной корреляции $\rho^2_{j\cdot 1,2,\ldots,m}$ называется *множественным коэффициентом детерминации.* Коэффициенты множественной корреляции и детерминации — величины положительные, принимающие значения в интервале $0<\rho_{j\cdot 1,2,\ldots,m}<1$. Оценками этих коэффициентов являются выборочные множественные коэффициенты корреляции и детерминации, которые обозначают соответственно $R_{j\cdot 1,2,\ldots,m}$ и $R^2_{j\cdot 1,2,\ldots,m}$. Формула для вычисления выборочного множественного коэффициента корреляции имеет вид

$$R_{j\cdot 1,2,\ldots,m} = \sqrt{1-|q_m|/q_{jj}}, \qquad (9.21)$$

где $|q_m|$ — определитель корреляционной матрицы, составленной из выборочных парных коэффициентов корреляции; q_{jj} — алгебраическое дополнение к элементу r_{jj}.

Многомерный корреляционный анализ позволяет получить оценку функции регрессии — уравнение регрессии. Коэффициенты в уравнении регрессии можно найти непосредственно через выборочные парные коэффици-

ситы корреляции или воспользоваться методом многомерной регрессии (см. § 10.7). В этом случае все предпосылки регрессионного анализа оказываются выполненными и, кроме того, связь между переменными строго линейна.

Пример 9.5. Вычислить оценки множественных коэффициентов корреляции и детерминации первого признака со всеми остальными, используя данные примера 9.4.

Решение. Оценку множественного коэффициента корреляции найдем по формуле

$$R_{1.2345} = \sqrt{1 - |q_4| / q_{11}}.$$

Вычислив определитель корреляционной матрицы, найдем $|q_4| = 0{,}11$. Алгебраическое дополнение к элементу r_{11} вычислено в примере 9.4 и равно $q_{11} = 0{,}56$. Тогда $R_{1.2345} = 0{,}89$; $R^2_{1.2345} = 0{,}80$.

Значимость множественного коэффициента корреляции и детерминации проверяют с помощью F-критерия Фишера. Проверяется нулевая гипотеза $H_0: \rho^2 = 0$. Для этого вычисляют статистику $F = \frac{R^2(n-l-2)}{l(1-R^2)}$, имеющую F-распределение с $k_1 = n-l-2$ и $k_2 = l$ степенями свободы. По табл. 4 приложений для уровня значимости α находят F_{α, k_1, k_2}. Если $F > F_{\alpha, k_1, k_2}$, нулевая гипотеза отвергается. Коэффициент считают значимым. При $F < F_{\alpha, k_1, k_2}$ нет оснований отвергать нулевую гипотезу.

§ 9.9. Ранговая корреляция

В некоторых случаях встречаются признаки, не поддающиеся количественной оценке (назовем такие признаки *объектами*). Попытаемся, например, оценить соотношение между математическими и музыкальными способностями группы учащихся. «Уровень способностей» является переменной величиной в том смысле, что он варьирует от одного индивидуума к другому. Его можно измерить, если выставлять каждому индивидууму отметки. Однако этот способ лишен объективности, так как разные экзаменаторы могут выставить одному и тому же учащемуся разные отметки. Элемент субъективизма можно исключить, если учащиеся будут ранжированы. Расположим учащихся по порядку, в соответствии со степенью способностей и присвоим каждому из них порядковый номер, который назовем *рангом*. Корреляция между рангами более точно отражает соотношение между способностями учащихся, чем корреляция между отметками.

Тесноту связи между рангами измеряют так же, как и между признаками. Рассмотрим уже известную формулу коэффициента корреляции

$$r = \frac{\sum_{x,y}(x-\bar{x})(y-\bar{y})}{ns_x s_y}.$$

Пусть $(x-\bar{x})=x'$, $(y-\bar{y})=y'$; тогда, учитывая, что $s_x=\sqrt{\sum_x (x-\bar{x})^2/n}$, $s_y=\sqrt{\sum_y (y-\bar{y})^2/n}$, можно записать

$$r = \left(\sum_{x,y} x'y'\right) \Big/ \sqrt{\sum_x x'^2 \sum_y y'^2}. \qquad (9.22)$$

В зависимости от того, что принять за меру различия между величинами x', y', можно получить различные коэффициенты связи между рангами. Обычно используют коэффициент корреляции рангов Кэнделла (τ) и коэффициент корреляции рангов Спирмэна ρ.

Введем следующую меру различия между объектами: будем считать $x'=+1$, если $i<j$, и $x'=-1$, если $i>j$. Соответственно $y'=+1$, если $i<j$, и $y=-1$, если $i>j$. Поясним сказанное на примере. Имеем две последовательности:

X	2	4	5	1	3
Y	1	5	3	4	2

Рассмотрим отдельно каждую из них. В последовательности X первой паре элементов — 2; 4 припишем значение $+1$, так как $i<j$; второй паре 2; 5 также припишем значение $+1$, третьей паре 2; 1 припишем значение -1, поскольку $i>j$, и т.д. Последовательно перебираем все пары, причем каждая пара должна быть учтена один раз. Так, если учтена пара 2; 1, то не следует учитывать пару 1; 2. Аналогичные действия проделаем с последовательностью Y, причем порядок перебора пар должен в точности повторять порядок перебора пар в последовательности X. Результаты этих действий представим в виде табл. 9.3.

Рассмотрим формулу (9.22). В нашем случае $\Sigma x'^2 = \Sigma y'^2$ и равна количеству пар, участвовавших в пере-

Таблица 9.3

X	x'	Y	y'	$x'y'$
2;4	+1	1;5	+1	+1
2;5	+1	1;3	+1	+1
2;1	−1	1;4	+1	−1
2;3	+1	1;2	+1	+1
4;5	+1	5;3	−1	−1
4;1	−1	5;4	−1	+1
4;3	−1	5;2	−1	+1
5;1	−1	3;4	+1	−1
5;3	−1	3;2	−1	+1
1;3	+1	4;2	−1	−1
Σ				+2

боре. Каждая пара встречается только один раз, поэтому их общее количество равно числу сочетаний из n по 2, т. е. $C_n^2=n(n-1)/(1\cdot 2)$. Обозначая $\Sigma x'y'=s$, получаем формулу коэффициента корреляции рангов Кэнделла:

$$\tau=\frac{s}{0{,}5n(n-1)}=\frac{2s}{n(n-1)}. \qquad (9.23)$$

Теперь рассмотрим другую меру различия между объектами. Если обозначить через $\bar{X}$ средний ранг последовательности X, через $\bar{Y}$ — средний ранг последовательности Y, то $x'=X-\bar{X}$, $y'=Y-\bar{Y}$. Поскольку ранги последовательности X и Y есть числа натурального ряда, то их сумма равна $^1/_2\ n(n+1)$, а средний ранг $\bar{X}=\bar{Y}=(n+1)/2$. Тогда $\Sigma x'^2=\sum\left(X-\frac{n+1}{2}\right)^2=\Sigma X-(n+1)\Sigma X+\frac{n(n+1)^2}{4}$. Сумма квадратов чисел натурального ряда равна $n(n+1)(2n+1)/6$. Тогда

$$\Sigma x'^2=n(n+1)(2n+1)/6-n(n+1)^2/2+n(n+1)^2/4=$$
$$=n(n+1)[(2n+1)/6-(n+1)/4]=(n^3-n)/12;$$
$$\Sigma y'^2=\Sigma x'^2=(n^3-n)/12.$$

Введем новую величину d, равную разности между рангами: $d=X-Y$, и определим через нее величину $\Sigma x'y'$. Имеем:

$$\Sigma d^2 = \Sigma (X - Y)^2 = \Sigma (x' - y')^2 = \Sigma x'^2 + \Sigma y'^2 - 2\Sigma x' y'; \quad \Sigma x' y' = [(n^3 - n)/6 - \Sigma d^2]/2.$$

Коэффициент корреляции рангов Спирмэна

$$\rho = 1 - 6\Sigma d^2/(n^3 - n). \qquad (9.24)$$

У коэффициентов τ и ρ разные масштабы, они отличаются шкалами измерений. Поэтому на практике нельзя ожидать, что они совпадут. Чаще всего, если значения обоих коэффициентов не слишком близки к 1, ρ по абсолютной величине примерно на 50% превышает τ. Выведены неравенства, связывающие ρ и τ. Например, при больших n можно пользоваться следующим приближенным соотношением: $-1 \leqslant 3\tau - 2\rho \leqslant 1$, или $^1/_2\tau^2 + \tau - ^1/_2 \leqslant \rho \leqslant ^3/_2\tau + ^1/_2$. Коэффициент ρ легче рассчитать, однако с теоретической точки зрения больший интерес представляет коэффициент τ.

При вычислении коэффициента корреляции рангов Кэнделла для подсчета s можно использовать следующий прием: одну из последовательностей упорядочивают так, чтобы ее элементы были числами натурального ряда; соответственно изменяют и другую последовательность. Тогда сумму $\Sigma x'y'$ можно подсчитывать лишь по последовательности Y, так как все x' равны $+1$.

Пример 9.6. Установить, имеется ли корреляционная связь между интенсивностью окраски пряжи в 10 партиях сырья, предназначенного для текстильной промышленности, и его влажностью. Вычислить коэффициенты корреляции ρ и τ. Эксперт расположил партии сырья в следующем порядке (табл. 9.4).

Таблица 9.4

Интенсивность окраски X	3	8	5	4	2	10	1	7	9	6
Влажность Y	1	9	10	2	4	7	3	5	8	6
d	2	−1	−5	2	−2	3	−2	2	1	0
d^2	4	1	25	4	4	9	4	4	1	0

Решение. Вычислим по формуле (9.24) коэффициент корреляции рангов Спирмэна ρ. Для этого подсчитаем d и d^2. Это удобно сделать непосредственно в исходной таблице. Имеем: $\Sigma d^2 = 56$; $\rho = 1 - 6 \cdot 56/(1000 - 10) = 0{,}66$.

Вычислим по формуле (9.23) коэффициент корреляции рангов Кэнделла τ. Для этого упорядочим последовательность X и соответственно преобразуем последовательность Y. Имеем:

X	1	2	3	4	5	6	7	8	9	10
Y	3	4	1	2	10	6	5	9	8	7

Теперь $\Sigma x'y' = s$ можно найти лишь по последовательности Y, так как все x' имеют значения $+1$, поскольку все $i < j$. Для подсчета s составим табл. 9.5.

Таблица 9.5

+1	−1	−1	+1	+1	+1	+1	+1	+1	+5
	−1	−1	+1	+1	+1	+1	+1	+1	+4
		+1	+1	+1	+1	+1	+1	+1	+7
			+1	+1	+1	+1	+1	+1	+6
				−1	−1	−1	−1	−1	−5
					−1	+1	+1	+1	+2
						+1	+1	+1	+3
							−1	−1	−2
								−1	−1
									+19

Расхождение в оценке тесноты связи между двумя коэффициентами довольно значительное (этот пример показывает, насколько проще вычислять коэффициент Спирмэна).

Итак, $\Sigma x'y' = s = +19$; $\tau = 2 \cdot 19/10 \cdot 9 = 38/90 = 0{,}42$.

Установлено, что распределение ρ и τ при $n \to \infty$ приближается к нормальному, поэтому довольно просто оценить значимость этих коэффициентов (см. § 9.6).

Если нельзя установить ранговое различие нескольких объектов, говорят, что такие объекты являются *связанными*. В этом случае объектам приписывается средний ранг. Например, если связанными являются объекты 4 и 5, то им приписывают ранг 4.5; если связанными являются объекты 1, 2, 3, 4 и 5, то их средний ранг $(1+2+3+4+5)/5 = 3$. Сумма рангов связанных объектов должна быть равна сумме рангов при ранжировании без связей. Формулы коэффициентов корреляции для ρ и τ в этом случае также можно вывести из формулы обобщенного коэффициента корреляции, только знаменатель выражения (9.21) в этом случае не равен $n(n-1)/2$. Если t последовательных членов связаны, то все оценки, относящиеся к любой выбранной из них паре, равны нулю; число таких пар $t(t-1)$. Следователь-

но, $\Sigma x'^2 = [n(n-1) - \sum_t t(t-1)]/2$. Соответственно для другой последовательности $\Sigma y'^2 = [n(n-1) - \sum_u u(u-1)]/2$, где t и u — число связанных пар в последовательностях.

Обозначая $T = \sum_t t(t-1)/2$, $U = \sum_u u(u-1)/2$, получаем

$$\tau = \frac{s}{\sqrt{n(n-1)/2 - T}\sqrt{n(n-1)/2 - U}}.$$

Аналогично находим выражение для ρ. Только в этом случае $T = \sum_t (t^3-t)/12$, $U = \sum_u (u^3-u)/12$, где t и u — число связанных пар в последовательностях, а

$$\rho = \frac{(n^3-n)/6 - (T+U) - \Sigma d^2}{\sqrt{(n^3-n)/6 - 2T}\sqrt{(n^3-n)/6 - 2U}}.$$

Пример 9.7. На соревнованиях по фигурному катанию судьи следующим образом расположили участников соревнований:

Участники	*А*	*Б*	*В*	*Г*	*Д*	*Е*	*Ж*	*З*	*И*	*К*
I судья	1,5	1,5	3	4	6	6	6	8	9,5	9,5
II судья	1	2	4	4	4	6	7	8	9	10

Установить, насколько объективны оценки судей, т. е. насколько тесна связь между оценками. Вычислить коэффициенты ранговой корреляции ρ и τ.

Р е ш е н и е. Сначала вычислим коэффициент ρ. Для этого необходимо найти T и U. Первый судья поделил первое место между участниками *А* и *В*. Их объединенный ранг $(1+2)/2 = 1{,}5$. Участники *Д*, *Е*, *Ж* поделили 5, 6 и 7-е места. Их объединенный ранг равен 6 и т. д. При вычислении T имеем: *А* и *В* — два объединенных ранга, *Д*, *Е*, *Ж* — три объединенных ранга и *И*, *К* — два объединенных ранга. Таким образом,

$$T = [(2^3-2) + (3^3-3) + (2^3-2)]/12 = 3, \qquad U = (3^3-3)/12 = 2;$$

$$\rho = \frac{(1000-10)/6 - (3+2) - 7}{\sqrt{[1000-10)/6 - 6]\,[(1000-10)/6 - 4]}} = 0{,}956.$$

Для вычисления коэффициента корреляции рангов τ необходимо найти значение s, т. е. $\Sigma x'y'$. Если $i=j$, то $x'=0$ (или $y'=0$). Составим таблицу для подсчета $\Sigma x'y'$ (табл. 9.6).

Таблица 9.6

$x^1y^1x^1y^1$	$x^1y^1x^1y^0$	$x^1y^1x^0y^1$	$x^1y^1x^0y^0$	$x^1y^0x^1y^1$	$x^1y^0x^1y^0$	$x^1y^0x^0y^1$	$x^1y^0x^0y^0$
0 1 0	1 1 1	1 0 0	1 0 0	0 1 0	0 1 0	1 1 1	1 1 1
1 1 1	1 1 1	1 1 0	0 1 1	1 0 0	1 1 1	1 1 1	1 1 1
1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	Σ 2
1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	Σ 3	
1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	Σ 3		
1 1 1	1 1 1	1 1 1	1 1 1	Σ 3			
1 1 1	1 1 1	1 1 1	Σ 5				
1 1 1	1 1 1	Σ 5					
1 1 1	Σ 8						
Σ 8							

0 1 0
Σ 0

Порядок заполнения таблицы описан выше.
Далее находим:

$$s = 8 + 8 + 5 + 5 + 3 + 3 + 3 + 2 + 0 = 37,$$
$$T = (2 \cdot 1 + 3 \cdot 2 + 2 \cdot 1)/2 = 5, \quad U = (3 \cdot 2)/2 = 3.$$

Отсюда $\tau = 37/\sqrt{(45-5)(45-3)} = 37/\sqrt{40 \cdot 42} = 0{,}902$. Итак, можно сделать вывод, что оценки, данные судьями участникам соревнований, в достаточной мере объктивны. Это подтверждают величины обоих коэффициентов корреляции.

Если имеется несколько последовательностей, то возникает необходимость определить общую меру согласованности между ними. Такой мерой является *коэффициент конкордации.*

Пусть m — число последовательностей, n — количество рангов в каждой последовательности. Тогда коэффициент конкордации

$$W = \frac{12\Sigma d^2}{m^2(n^3-n)}, \tag{9.25}$$

где d — фактически встречающееся отклонение от среднего значения суммы рангов одного объекта.

Пример 9.8. Судьи расставили 6 участников соревнований по фигурному катанию следующим образом:

Участники	А	Б	В	Р	Д	Е
Судья I	5	4	1	6	3	2
Судья II	4	6	2	3	5	1
Судья III	3	4	2	6	5	1
Сумма рангов	12	14	5	15	13	4

Установить, насколько согласованы действия судей.

Р е ш е н и е. Для вычисления коэффициента конкордации необходимо найти d. Предварительно подсчитаем среднее значение суммы рангов одного объекта. Это можно сделать, вычислив сумму рангов, выставленную тремя судьями по всем объектам и разделив ее на число объектов, т. е. $63/6=10{,}5$, или же можно воспользоваться формулой $m(n+1)/2=3\cdot 7/2=10{,}5$; d получаем как разность между суммой рангов, выставленных одним судьей, и средней суммой рангов. Имеем.

d	1,5	3,5	−5,5	4,5	2,5	−6,5
d^2	2,25	12,25	30,25	20,25	6,25	42,25

$$\Sigma d^2 = 113{,}5, \quad W = 12113{,}5/9(216-6) = 0{,}72.$$

Величина коэффициента корреляции говорит о том, что действия судей хорошо согласованы.

Коэффициент корреляции рангов может быть использован для быстрого оценивания взаимосвязи между признаками, не имеющими нормального распределения, и полезен в тех случаях, когда признаки поддаются ранжированию, но не могут быть точно измерены.

Глава 10

Регрессионный анализ

§ 10.1. Основные положения регрессионного анализа

Основная задача регрессионного анализа — изучение зависимости между результативным признаком Y и наблюдавшимся признаком X, оценка функции регрессии.

Предпосылки регрессионного анализа:

1) Y — независимые случайные величины, имеющие постоянную дисперсию;

2) X — величины наблюдаемого признака (величины не случайные);

3) условное математическое ожидание $M(Y|X=x)$ можно представить в виде

$$M(Y|X=x)=\varphi(x)=\beta_0+\beta_1 x. \qquad (10.1)$$

Выражение (10.1), как уже упоминалось в § 9.2, называется функцией регрессии (или модельным уравнением регрессии) Y на X. Оценке в этом выражении подлежат параметры β_0 и β_1, называемые *коэффициентами* регрессии, а также $\sigma^2_{ост}$ — *остаточная дисперсия.*

Остаточной дисперсией называется та часть рассеивания результативного признака, которую нельзя объяснить действием наблюдаемого признака. Остаточная дисперсия может служить для оценки точности подбора вида функции регрессии (модельного уравнения регрессии), полноты набора признаков, включенных в анализ. Оценки параметров функции регрессии находят, используя метод наименьших квадратов.

В данной главе рассмотрен линейный регрессионный анализ. Линейным он называется потому, что изучает лишь те виды зависимостей $\varphi(x)$, которые линейны по оцениваемым параметрам, хотя могут быть нелинейны

по переменным X. Например, зависимости $M(Y|X=x)=\varphi_1(x)=\beta_0+\beta_1 x+\beta_2 x$, $M(Y|X=x)=\varphi_2(x)=\beta_0+\beta_1/x$, $M(Y|X=x)=\varphi_3(x)=\beta_0+\beta_1 x+\beta_2 x^2$ линейны относительно параметров β_0, β_1, β_2, хотя вторая и третья зависимости нелинейны относительно переменных x. Вид зависимости $\varphi(x)$ выбирают, исходя из визуальной оценки характера расположения точек на поле корреляции; опыта предыдущих исследований; соображений профессионального характера, основанных на знании физической сущности процесса.

Важное место в линейном регрессионном анализе занимает так называемая «нормальная регрессия». Она имеет место, если сделать предположения относительно закона распределения случайной величины Y. Предпосылки «нормальной регрессии»:

1) Y — независимые случайные величины, имеющие постоянную дисперсию и распределенные по нормальному закону;

2) X — величины наблюдаемого признака (величины не случайные);

3) условное математическое ожидание $M(Y|X=x)$ можно представить в виде (10.1).

В этом случае оценки коэффициентов регрессии — несмещенные с минимальной дисперсией и нормальным законом распределения. Из этого положения следует, что при «нормальной регрессии» имеется возможность оценить значимость оценок коэффициентов регрессии, а также построить доверительный интервал для коэффициентов регрессии и условного математического ожидания $M(Y|X=x)$.

§ 10.2. Линейная регрессия

Рассмотрим простейший случай регрессионного анализа — модель вида (10.1), когда зависимость $\varphi(x)$ линейна и по оцениваемым параметрам, и по переменным. Оценки параметров модели (10.1) β_0 и β_1 обозначим b_0 и b_1. Оценку остаточной дисперсии $\sigma^2_{ост}$ обозначим $s^2_{ост}$. Подставив в формулу (10.1) вместо параметров их оценки, получим уравнение регрессии $\bar{y}(x)=b_0+b_1 x_1$, коэффициенты которого b_0 и b_1 находят из условия минимума суммы квадратов отклонений измеренных зна-

чений результативного признака y_i от вычисленных по уравнению регрессии $y(x_i)$:

$$Q = \sum_{i=1}^{n} [y_i - y(x_i)]^2 = \min, \text{ или}$$

$$Q = \sum_{i=1}^{n} [y_i - b_0 - b_1 x_i]^2 = \min.$$

Составим систему нормальных уравнений: первое уравнение

$$\frac{\partial Q}{\partial b_0} = 2 \sum_{i=1}^{n} [y_i - b_0 - b_1 x_i] = 0,$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a_0 - \sum_{i=1}^{n} b_1 x_i = 0,$$

откуда $nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$;

второе уравнение

$$\frac{\partial Q}{\partial b_1} = 2 \sum_{i=1}^{n} x_i [y_i - b_0 - b_1 x_i] = 0,$$

откуда $b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$.

Итак,

$$\begin{cases} nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \end{cases} \qquad (10.2)$$

Оценки, полученные по способу наименьших квадратов, обладают минимальной дисперсией в классе линейных оценок. Решая систему (10.2) относительно b_0 и b_1, найдем оценки параметров β_0 и β_1:

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n}, \qquad (10.3)$$

$$b_0 = \left(\sum_{i=1}^{n} y_i\right) \Big/ n - b_1 \left(\sum_{i=1}^{n} x_i\right) \Big/ n = \bar{y} - b_1 \bar{x}. \qquad (10.4)$$

Остается получить оценку параметра $\sigma^2_{\text{ост}}$. Имеем

$$s^2_{\text{ост}} = \frac{1}{n-2} \sum_{i=1}^{n} [y_i - y(x_i)]^2, \qquad (10.5)$$

где n — количество наблюдений.

Если n велико, то для упрощения расчетов наблюдавшиеся данные принято группировать, т. е. строить корреляционную таблицу. Пример построения такой таблицы приведен в § 9.5. Формулы для нахождения коэффициентов регрессии по сгруппированным данным те же, что и для расчета по несгруппированным данным, но суммы $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i y_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i^2$ заменяют на

$$\sum_x x m_x, \sum_y y m_y, \sum_{xy} xy m_{xy}, \sum_x x^2 m_x, \sum_y y^2 m_y,$$

где m_x, m_y и m_{xy} — частоты повторений соответствующих значений переменных. В дальнейшем часто используется этот наглядный прием вычислений.

Пример 10.1. По данным, приведенным в табл. 9.1, оценить параметры уравнения регрессии. Предполагается, что связь выражается формулой (10.1).

Решение. При сгруппированных данных удобно использовать следующие формулы:

$$b_1 = \frac{\sum_{x,y} xy m_{xy} - \left(\sum_x x m_x \sum_y y m_y\right) \Big/ \sum_{xy} m_{xy}}{\sum_x x^2 m_x - \left(\sum_x x m_x\right)^2 \Big/ \sum_{xy} m_{xy}},$$

$$b_0 = \sum_y y m_y \Big/ \sum_y m_y - b_1 \sum_x x m_x \Big/ \sum_x m_x.$$

Используя суммы, вычисленные в табл. 9.2, получаем: $b_1 = -5{,}167$; $b_0 = 22{,}082$. Уравнение регрессии имеет вид $\bar{y}(x) = 22{,}082 + 5{,}167\,x$.

Таблица 10.1

Объем выполненных работ y, млн. руб. \ Накладные расходы x, млн. руб.	1—2 1,5	2—3 2,5	3—4 3,5	4—5 4,5	5—6 5,5	6—7 6,5	7—8 7,5	8—9 8,5	m_y	ym_y	y^2m_y	Σ^*yxm_{xy}
10—20 15	4	5							9	135	2 025	277,5
20—30 25	1	3	1						5	125	3 125	312,5
30—40 35	2	3	6	5	3	1			20	700	24 500	2695,0
40—50 45	1	5	9	19	8	7	2	1	51	2295	103 275	10912,5
50—60 55		1	2	7	16	9	4	2	41	2255	124 025	12897,5
60—70 65			1	5	6	4	2	2	20	1300	84 500	7605,0
70—80 75							1	3	4	300	22 500	2475,0
m_x	7	17	19	36	33	21	9	8	150	7110	363 950	37175,0
xm_x	10,5	42,5	66,5	162,0	181,5	136,5	67,5	68,0	735,0			
x^2m_x	15,75	106,25	232,75	729,0	998,25	887,25	506,25	578,0	4053,5			

Таблица 10.2

y_x	$y(x_l)$	m_x	$y_l - y(x_l)$	$[y_l - y(x_l)]^2 m_x$	$y - \bar{y}$	$(y - \bar{y})^2 m_x$	$(x_l - \bar{x})^2 m_x$
22,14	29,832	7	—7,69	414,952	—25,26	4466,473	80,92
31,47	35,00	17	—3,53	211,835	—15,93	4314,003	97,92
42,89	40,166	19	2,72	140,982	—4,51	386,462	37,24
48,33	45,336	36	3,00	322,704	0,93	31,136	5,78
52,57	50,500	33	2,07	141,402	5,17	882,054	11,88
52,62	55,668	21	—3,05	195,196	5,12	550,502	53,76
57,22	60,834	9	—3,61	117,548	9,82	867,892	60,84
63,75	66,002	8	—2,25	40,554	15,35	1884,980	103,68
Σ		150		1585,073		13383,502	452,02

Промежуточные результаты вычислений оценки остаточной дисперсии приведены в табл. 10.2. По данным табл. 10.1 находим: $\bar{y} = 7110/150 = 47{,}4$, $\bar{x} = 735/150 = 4{,}9$.

Для сгруппированного ряда остаточную дисперсию оценим по формуле

$$s^2_{\text{ост}} = \left[\sum_x [y_l - y(x_l)]^2 m_x\right]/(n-2),$$

откуда $s^2_{\text{ост}} = 1585{,}073/148 = 10{,}711$, $s_{\text{ост}} = 3{,}38$.

§ 10.3. Нелинейная регрессия

Рассмотрим случай, когда зависимость нелинейна по переменным x, например модель вида

$$M(Y \mid X = x) = \varphi(x) = \beta_0 + \beta_1/x. \qquad (10.6)$$

На рис. 10.1 изображено поле корреляции. Очевидно, что зависимость между Y и X нелинейная и ее графическим изображением является не прямая, а кривая. Оценкой выражения (10.6) является уравнение регрессии

$$\bar{y}(x) = b_0 + b_1/x,$$

где b_0 и b_1 — оценки коэффициентов регрессии β_0 и β_1.

Принцип нахождения коэффициентов тот же — метод наименьших квадратов, т. е.

$$Q = \sum_{i=1}^{n} [y_i - y(x_i)]^2 = Q_{\min},$$

или

$$Q = \sum_{i=1}^{n} [y_i - b_0 - b_1/x_i]^2 = Q_{\min}.$$

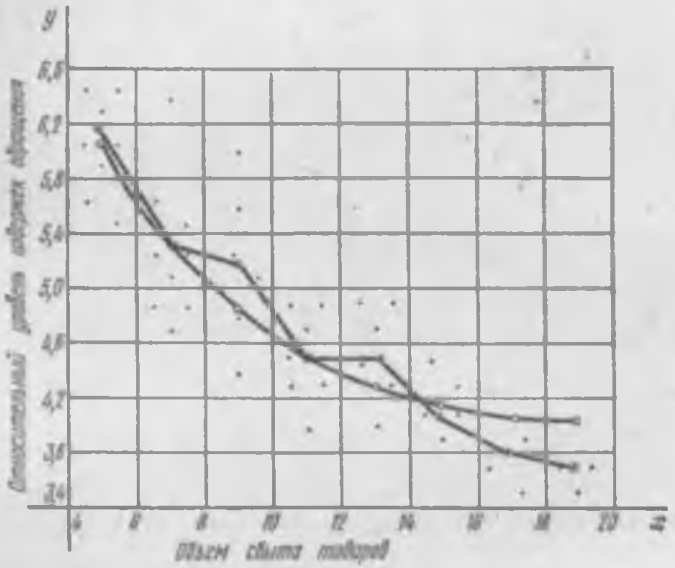

Рис. 10.1

Дифференцируя последнее равенство по b_0 и b_1 и приравнивая правые части нулю, получаем так называемую систему нормальных уравнений:

$$\begin{cases} \sum_{i=1}^{n} b_0 + \sum_{i=1}^{n} b_1/x_i = \sum_{i=1}^{n} y_i, \\ \sum_{i=1}^{n} b_0/x_i + \sum_{i=1}^{n} b_1/x_i^2 = \sum_{i=1}^{n} y_i/x_i. \end{cases} \qquad (10.7)$$

Пример 10.2. По данным, приведенным в табл. 10.3, вычислить коэффициенты в уравнении регрессии. Связь между переменными имеет вид (10.6).

Таблица 10.3

Относительный уровень издержек обращения y, % \ Объем сбыта товаров x, млн. руб.	4—6 5	6—8 7	8—10 9	10—12 11	12—14 13	14—16 15	16—18 17	18—20 19	m_y	ym_y	y^2m_y	ym_{xy}/x
3,4—3,8 3,6							2	3	5	18	64,8	0,9918
3,8—4,2 4,0				1	1	3	2	1	8	32	128	2,1524
4,2—4,6 4,4			1	4	2	2			9	39,6	174,24	3,3523
4,6—5,0 4,8		3	1	3	3				10	48	230,4	5,0074
5,0—5,4 5,2		3	2						5	26	135,2	3,3842
5,4—5,8 5,6	2	2	2						5	28	156,8	4,4621
5,8—6,2 6,0	3		1						4	24	144	4,2666
6,2—6,6 6,4	3	1							4	25,6	163,84	4,7546
m_x	8	9	6	8	6	5	4	4	50	241,2	1197,28	28,3714
$\frac{1}{x} m_x$	1,600	1,2857	0,6667	0,7273	0,4615	0,3333	0,2353	0,2105	5,5203			
$\frac{1}{x^2} m_x$	0,3200	0,1837	0,0741	0,0661	0,0355	0,0222	0,0138	0,0111	0,7265			
$\overline{y}_{i x=} \frac{\Sigma y m_{xy}}{m_x}$	6,05	5,29	5,20	4,50	4,53	4,16	3,80	3,70				

Р е ш е н и е. В табл. 10.3 приведены все необходимые суммы для составления нормальных уравнений, а также средние значения $\bar{y}_i(x)$ для каждого интервала изменения x. Значения $\bar{y}_i(x)$ приведены в последней строке таблицы и нанесены на поле корреляции (рис. 10.1). Решив уравнения (10.7), можно вычислить значения $y(x_i)$ и построить линию регрессии. Запишем систему уравнений (10.7) в виде

$$\begin{cases} nb_0 + b_1 \sum\limits_x m_x/x = \sum\limits_y y m_y, \\ b_0 \sum\limits_x m_x/x + b_1 \sum\limits_x m_x/x^2 = \sum\limits_{xy} y m_{xy}/x, \end{cases}$$

или

$$\begin{cases} 50b_0 + 5{,}5203b_1 = 241{,}2, \\ 5{,}5203b_0 + 0{,}7265b_1 = 28{,}3714, \end{cases}$$

откуда $b_1 = 14{,}881$, $b_0 = 3{,}181$. Уравнение регрессии (10.6) теперь принимает вид $\bar{y}(x) = 3{,}181 + 14{,}881/x$.

Используя формулу для оценки дисперсии по данным сгруппированного ряда, вычислим $s^2_{ост}$ (табл. 10.4). Среднее значение $\bar{y} = \Sigma y m_y/n = 241{,}2/50 = 4{,}82$.

Таблица 10.4

y_i	$y(x_i)$	m	x	$y_i - y(x_i)$	$[y_i - y(x_i)]^2 m$	$(y_i - \bar{y})^2 m$
6,05	6,16	8	5	—0,11	0,0968	12,1032
5,29	5,31	9	7	—0,02	0,0036	1,9881
5,20	4,83	6	9	+0,37	0,8214	0,8664
4,50	4,53	8	11	—0,03	0,0072	0,8192
4,53	4,32	6	13	+0,21	0,2646	0,5046
4,16	4,17	5	15	—0,01	0,0005	2,1780
3,80	4,05	4	17	—0,25	0,2500	4,1616
3,70	3,96	4	19	—0,26	0,2704	5,0176
Σ					1,7145	27,6387

Остаточная дисперсия $s^2_{ост} = 1{,}7145/48 = 0{,}358$.

Полученная линия регрессии нанесена на рис. 10.1. Как следует из рисунка, форма связи (гипербола) выбрана правильно. Полученная линия регрессии является хорошей аппроксимацией эмпирических данных.

В общем случае нелинейной зависимости между переменными Y и X связь может выражаться многочленом k-й степени от x:

$$M(Y \mid X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k.$$

Коэффициенты регрессии определяют по принципу наименьших квадратов. Система нормальных уравнений имеет вид

$$\begin{cases} b_0 n + b_1 \sum\limits_{i=1}^{n} x_i + b_2 \sum\limits_{i=1}^{n} x_i^2 + \ldots + b_k \sum\limits_{i=1}^{n} x_i^k = \sum\limits_{i=1}^{n} y_i, \\ b_0 \sum\limits_{i=1}^{n} x_i + b_1 \sum\limits_{i=1}^{n} x_i^2 + b_2 \sum\limits_{i=1}^{n} x_i^3 + \ldots + b_k \sum\limits_{i=1}^{n} x_i^{k+1} = \sum\limits_{i=1}^{n} y_i x_i, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ b_0 \sum\limits_{i=1}^{n} x_i^k + b_1 \sum\limits_{i=1}^{n} x_i^{k+1} + b_2 \sum\limits_{i=1}^{n} x_i^{k+2} + \ldots + \\ + b_k \sum\limits_{i=1}^{n} x_i^{2k} = \sum\limits_{i=1}^{n} y_i x_i^k. \end{cases}$$

Вычислив коэффициенты системы, ее можно решить любым известным способом.

Пример 10.3. По данным, приведенным в табл. 10.5, оценить параметры уравнения регрессии. Переменные связаны между собой зависимостью $M(Y|X=x)=\beta_0+\beta_1 x+\beta_2 x^2$ (уравнение параболы).

Р е ш е н и е. Система нормальных уравнений имеет вид

$$\begin{cases} b_0 n + b_1 \sum\limits_{1}^{50} x_i + b_2 \sum\limits_{1}^{50} x_i^2 = \sum\limits_{1}^{50} y_i, \\ b_0 \sum\limits_{1}^{50} x_i + b_1 \sum\limits_{1}^{50} x_i^2 + b_2 \sum\limits_{1}^{50} x_i^3 = \sum\limits_{1}^{50} y_i x_i, \\ b_0 \sum\limits_{1}^{50} x_i^2 + b_1 \sum\limits_{1}^{50} x_i^3 + b_2 \sum\limits_{1}^{50} x_i^4 = \sum\limits_{1}^{50} y x_i^2. \end{cases}$$

Используя данные табл. 10.5, получаем следующую систему нормальных уравнений:

$$\begin{cases} 50b_0 + 125b_1 + 365b_2 = 151, \\ 125b_0 + 365b_1 + 1175b_2 = 405, \\ 365b_0 + 1175b_1 + 4025b_2 = 1237, \end{cases} \text{ или } \begin{cases} b_0 + 2{,}5b_1 + 7{,}3b_2 = 3{,}02, \\ b_0 + 2{,}92b_1 + 9{,}4b_2 = 3{,}24, \\ b_0 + 3{,}22b_1 + 11{,}03b_2 = 3{,}39. \end{cases}$$

Решая данную систему уравнений, найдем значения коэффициентов регрессии: $b_0=1{,}424$, $b_1=0{,}799$, $b_2=-0{,}055$. Уравнение регрессии имеет вид

$$\bar{y} = 1{,}424 + 0{,}799x - 0{,}055x^2.$$

Оценим остаточную дисперсию. Все вычисления сведены в табл. 10.6.

Таблица 10.5

y \ x	1	2	3	4	m_y	ym_y	yx, m_{xy}	$yx^2 m_{xy}$
5				1	1	5	20	80
4		3	6	6	15	60	192	648
3	3	8	6	3	20	60	147	411
2	5	4	3		12	24	44	96
1	2				2	3	2	2
m_x	10	15	15	10	50	151	405	1237
xm_x	10	30	45	40	125			
$x^2 m_x$	10	60	135	160	365			
$x^3 m_x$	10	120	405	640	1175			
$x^4 m_x$	10	240	1215	2560	4025			

Таблица 10.6

y_i	$y(x_i)$	m	x	$[y_i - y(x_i)]^2 m$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2 m$
2,100	2,168	10	1	0,046	—0,92	8,464
2,933	2,802	15	2	0,256	—0,087	0,1125
3,200	3,326	15	3	0,237	+0,180	0,486
3,800	3,740	10	4	0,036	+0,780	6,084
Σ				0,575		15,1465

Остаточная дисперсия

$$s^2_{\text{ост}} = \frac{\Sigma [(y_i - y(x_i)]^2 m_x}{n - m - 1} = 0{,}575/47 = 0{,}0122.$$

§ 10.4. Оценка значимости коэффициентов регрессии. Интервальная оценка коэффициентов регрессии

Проверить значимость оценок коэффициентов регрессии — значит установить, достаточна ли величина оценки для статистически обоснованного вывода о том,

что коэффициент регрессии отличен от нуля. Для этого проверяют гипотезу о равенстве нулю коэффициента регрессии, соблюдая предпосылки «нормальной регрессии». В этом случае вычисляемая для проверки нулевой гипотезы $H_0: \beta = 0$ статистика

$$t = |b/s_b| \tag{10.8}$$

имеет распределение Стьюдента с $k = n-2$ степенями свободы (b — оценка коэффициента регрессии, s_b — оценка среднего квадратического отклонения коэффициента регрессии, иначе стандартная ошибка оценки). По уровню значимости α и числу степеней свободы k находят по таблицам распределения Стьюдента (см. табл. 5 приложений) критическое значение $t_{\alpha,k}$, удовлетворяющее условию $P(|t| \geqslant t_{\alpha,k}) \geqslant \alpha$. Если $|t| \geqslant t_{\alpha,k}$, то нулевую гипотезу о равенстве нулю коэффициента регрессии отвергают, коэффициент считают значимым. При $|t| < t_{\alpha,k}$ нет оснований отвергать нулевую гипотезу.

Оценки среднего квадратического отклонения коэффициентов регрессии вычисляют по следующим формулам:

$$s_{b_0} = s_{\text{ост}}/\sqrt{n-2},\ s_{b_1} = \frac{s_{\text{ост}}}{s_x\sqrt{n-2}}, \tag{10.9}$$

где $s_x = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}$; $s^2_{\text{ост}}$ — оценка остаточной дисперсии, вычисляемая по формуле (10.5).

Доверительный интервал для значимых параметров строят по обычной схеме (см. гл. 5). Из условия

$$P(-t_{\alpha,k} < (\beta - b)/s_b < t_{\alpha,k}) = 1 - \alpha,$$

где α — уровень значимости, находим

$$b - t_{\alpha,k}s_b < \beta < b + t_{\alpha,k}s_b. \tag{10.10}$$

Пример 10.4. Проверить значимость оценок коэффициентов регрессии вычисленных в примере 10.1, считая, что предпосылки «нормальной регрессии» соблюдены. Если оценки значимы, то построить доверительные интервалы для параметров β_0 и β_1.

Решение. Оценку остаточной дисперсии вычислим по формуле $s^2_{\text{ост}} = \Sigma[y_i - y(x_i)]^2 m_x/n - 2$. Необходимые суммы квадратов приведены в табл. 10.2. Имеем: $s^2_{\text{ост}} = 1585{,}073/148 = 10{,}711$, $s_{\text{ост}} = 3{,}28$. Далее находим $s_{b_0} = s_{\text{ост}}/\sqrt{n-2} = 0{,}27$, отсюда $t_{b_0} = 81{,}87$. Вычислим теперь $s_{b_1} = s_{\text{ост}}/s_x\sqrt{n-2}$. Сумму квадратов,

необходимую для определения s_{x_i}, возьмем из табл. 10.2: $s_x = \sqrt{452/150} = 1{,}77$; $s_{b_1} = 3{,}28/1{,}77\sqrt{148} = 0{,}15$. Вычислим статистику: $t_{b_1} = 5{,}17/0{,}15 = 34{,}45$. По табл. 5 приложений для $\alpha = 0{,}05$ и $k = 148$, находим $t_{0{,}05;\,148} = 196$. Итак, $87{,}78 > 1{,}96$, $34{,}46 > 1{,}96$. Следовательно, нулевую гипотезу о равенстве нулю коэффициента регрессии следует отвергнуть. Коэффициенты регрессии значимы.

Доверительные интервалы для коэффициентов регрессии таковы:

$$22{,}082 - 1{,}96 \cdot 0{,}27 < \beta_0 < 22{,}082 + 1{,}96 \cdot 0{,}27,$$
$$21{,}55 < \beta_0 < 22{,}61,$$
$$5{,}17 - 0{,}96 \cdot 0{,}15 < \beta_1 < 5{,}17 + 1{,}96 \cdot 0{,}15,$$
$$4{,}88 < \beta_1 < 5{,}46.$$

С вероятностью $P = 1 - \alpha = 0{,}95$ параметры не выйдут за вычисленные границы.

§ 10.5. Интервальная оценка для условного математического ожидания

Линия регрессии характеризует изменение условного математического ожидания результативного признака от вариации остальных признаков. Точечной оценкой условного математического ожидания $M(Y|X=x)$ является условное среднее $\bar{y}(x)$. Кроме точечной оценки для $M(Y|X=x)$ можно построить доверительный интервал в точке $x = x_0$.

Известно, что $(\bar{y}(x) - M(Y|X=x))/s_{\bar{y}(x)}$ имеет распределение Стьюдента с $k = n-2$ степенями свободы. Найдя оценку среднего квадратического отклонения для условного среднего, можно построить доверительный интервал для условного математического ожидания $M(Y|X=x)$ в точке $x = x_0$.

Оценку дисперсии условного среднего вычисляют по формуле

$$s^2_{\bar{y}(x)} = s^2_{\text{ост}} = \left[1/n + (x_0 - \bar{x})^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2\right],$$

или для интервального ряда

$$s^2_{\bar{y}(x)} = s^2_{\text{ост}} \left[1/n + (x_0 - \bar{x})^2 / \sum_{x} (x - \bar{x})^2 m_x\right]. \quad (10.11)$$

Доверительный интервал находят из условия

$$P\left(-t_{\alpha,k} < \frac{\bar{y}(x) - M(Y \mid X = x)}{s_{\bar{y}(x)}} < t_{\alpha,k}\right) = \alpha,$$

где α — уровень значимости. Отсюда

$$\bar{y}(x) - t_{\alpha,k}\, s_{\bar{y}(x)} < M(Y \mid X = x) < \bar{y}(x) + t_{\alpha,k}\, s_{\bar{y}(x)}.$$

Доверительный интервал для условного математического ожидания можно изобразить графически (рис. 10.2). Из рис. 10.2 видно, что в точке $x_0 = \bar{x}$ границы интервала наиболее близки друг другу. Расположение границ доверительного интервала показывает, что прогнозы по уравнению регрессии хороши только в случае, если значение x не выходит за пределы выборки, по которой вычислено уравнение регрессии; иными словами, экстраполяция по уравнению регрессии может привести к значительным погрешностям.

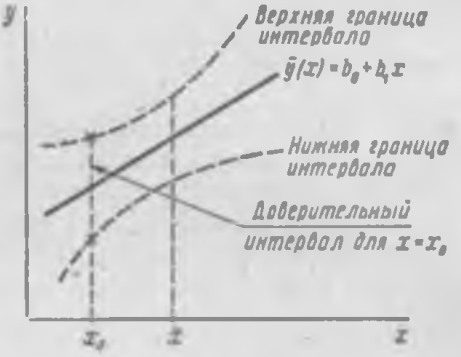

Рис. 10.2

Пример 10.5. Рассчитать доверительные интервалы для условного математического ожидания в точках $x_0 = 1{,}5$ и $x_0 = 4{,}5$, используя данные примера 10.1.

Решение. Сначала вычислим $\bar{y}(x_0)$, используя уравнение регрессии, полученное в примере 10.1. Промежуточные вычисления представим в виде табл. 10.7.

Таблица 10.7

x_0	$\bar{y}(x)$	$s_{\bar{y}(x_0)}$	$t_{\alpha,k} \cdot s_{\bar{y}(x_0)}$	$t_н$	$t_в$
1,5	29,83	0,59	1,15	28,68	30,98
4,5	45,34	0,27	0,53	44,81	45,87

Находим $s^2_{\bar{y}(x_0)}$. Имеем

$$s^2_{\bar{y}(1,5)} = 10{,}711\,[1/150 + (1{,}5 - 4{,}9)^2]/452 = 0{,}34, \quad s_{\bar{y}(1,5)} = 0{,}59.$$

Аналогично проводим вычисления для $x_0 = 4{,}5$. Затем по табл. 5 приложений для уровня значимости $\alpha = 0{,}05$ и $k = 150 - 2 = 148$ на-

ходим $t_{0,05;\,148}=1{,}96$. Определяем верхнюю и нижнюю границы интервала:

$$29{,}83-1{,}15<M(Y\mid X=1{,}5)<29{,}83+1{,}15,$$

$$45{,}34-0{,}53<M(Y\mid X=4{,}5)<45{,}34+0{,}53.$$

Таким образом, на краях границ ширина интервала больше, чем в середине (в точке, близкой к $\bar{x}=4{,}9$).

§ 10.6. Проверка значимости уравнения регрессии

Оценить значимость уравнения регрессии — значит установить, соответствует ли математическая модель, выражающая зависимость между Y и X, экспериментальным данным. Для оценки значимости в предпосылках «нормальной регресссии» проверяют гипотезу $H_0: \beta_1=0$. Если она отвергается, то считают, что между Y и X нет связи (или связь нелинейная). Для проверки нулевой гипотезы используют основное положение дисперсионного анализа о разбиении суммы квдратов на слагаемые. Воспользуемся разложением (8.3). Общая сумма квадратов отклонений результативного признака $Q=\sum_{i=1}^{n}(y_i-\bar{y})^2$ разлагается на Q_1 (сумму, характеризующую влияние признака X) и Q_2 (остаточную сумму квадратов, характеризующую влияние неучтенных факторов). Очевидно, чем меньше влияние неучтенных факторов, тем лучше математическая модель соответствует экспериментальным данным, так как вариация Y в основном объясняется влиянием признака X.

Для проверки нулевой гипотезы вычисляют статистику $F=(Q_1/Q_{ост})\cdot(k_2/k_1)$, которая имеет распределение Фишера — Снедекора с $k_1=1$, $k_2=n-2$ степенями свободы (n — число наблюдений). По уровню значимости α и числу степеней свободы k_1 и k_2 находят по таблицам F-распределение (см. табл. 4 приложений) критическое значение $F_{(\alpha,k_1,k_2)}$, удовлетворяющее условию $P(F>F_{(\alpha,k_1,k_2)})\geqslant\alpha$. Если $F>F_{(\alpha,k_1,k_2)}$, нулевую гипотезу отвергают, уравнение считают значимым. Если $F<F_{(\alpha,k_1,k_2)}$ то нет оснований отвергать нулевую гипотезу.

Пример 10.6. Оценить значимость уравнения регрессии, вычисленного в примере 10.1.

Р е ш е н и е. Для проверки нулевой гипотезы $H_0: \beta_1=0$ вычислим статистику: $F=(Q_1/Q_{ост})(k_2/k_1)$ при $n=150$, $k_1=1$, $k_2=148$,

$Q=13383{,}5$, $Q_{ост}=1585{,}07$; $Q_1=Q-Q_{ост}=11798{,}43$; статистика $F=(11798{,}43\cdot 148)/1585{,}07=1102{,}7$. По табл. 4 приложений при $k_1=1$, $k_2=148$ для $\alpha=0{,}05$ находим $F_{0{,}05;1;148}=3{,}84$. Имеем $F>F_{0{,}05;1;148}$, нулевую гипотезу отвергаем, линия регрессии значима, математическая модель хорошо согласуется с экспериментальными данными.

§ 10.7. Многомерный регрессионный анализ

В случае, если изменения результативного признака определяются действием совокупности других признаков, имеет место многомерный регрессионный анализ. Пусть результативный признак Y, а независимые признаки $(x_1, x_2, ..., x_m)$. Для многомерного случая предпосылки регрессионного анализа можно сформулировать следующим образом: Y — независимые случайные величины со средним $M(Y|X=x_1, x_2, ..., x_m)$ и постоянной дисперсией $\sigma^2_{ост}$; $x_1, x_2, ..., x_m$ — линейно независимые векторы $x(x_{11}, x_{12}, ..., x_{1n})$, ..., $x_m(x_{m1}, x_{m2}, ..., x_{mn})$. Все положения, изложенные в § 10.1, справедливы для многомерного случая. Рассмотрим модель вида

$$M(Y|X=x_1, x_2, \ldots, x_m)=\beta_0+\beta_1 x_1+\ldots+\beta_m x_m. \quad (10.13)$$

Оценке подлежат параметры β_0, β_1,..., β_m и остаточная дисперсия. Заменив параметры их оценками, запишем уравнение регрессии

$$\bar{y}(x)=b_0+b_1 x_1+\ldots+b_m x_m. \quad (10.14)$$

Коэффициенты в этом выражении находят методом наименьших квадратов.

Исходными данными для вычисления коэффициентов $b_0, b_1, ..., b_m$ является выборка из многомерной совокупности, представляемая обычно в виде матрицы X и вектора Y:

$$X=\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{j1} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{j2} & \cdots & x_{m2} \\ \cdots & \cdots & \cdots & & \cdots & \cdots \\ \cdots & \cdots & \cdots & x_{jl} & \cdots & \cdots \\ x_{1n} & x_{2n} & \cdots & x_{jn} & \cdots & x_{mn} \end{bmatrix}, \quad Y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}.$$

Как и в двумерном случае, составляют систему нормальных уравнений

$$\begin{cases} nb_0 + b_1 \sum_{i=1}^{n} x_{1i} + \ldots + b_m \sum_{i=1}^{n} x_{mi} = \sum_{i=1}^{n} y_i, \\ b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i} x_{1i} + \ldots + b_m \sum_{i=1}^{n} x_{1i} x_{mi} = \sum_{i=1}^{n} x_{1i} y_i, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ b_0 \sum_{i=1}^{n} x_{mi} + b_1 \sum_{i=1}^{n} x_{1i} x_{mi} + \ldots + b_m \sum_{i=1}^{n} x_{mi} x_{mi} = \sum_{i=1}^{n} x_{mi} y_i, \end{cases} \tag{10.15}$$

которую можно решить любым способом, известным из линейной алгебры. Рассмотрим один из них — способ обратной матрицы. Предварительно преобразуем систему уравнений. Выразим из первого уравнения значение b_0 через остальные параметры:

$$b_0 = \left(\sum_{i=1}^{n} y_i\right) \Big/ n - \left(b_1 \sum_{i=1}^{n} x_{1i}\right) \Big/ n - \ldots - \left(b_m \sum_{i=1}^{n} x_{mi}\right) \Big/ n = \bar{y} - b_1 \bar{x}_1 - \ldots - b_m \bar{x}_m. \tag{10.16}$$

Подставим в остальные уравнения системы вместо b_0 полученное выражение:

$$b_1 \left[\sum_{i=1}^{n} x_{1i}^2 - n(\bar{x}_1)^2\right] + b_2 \left(\sum_{i=1}^{n} x_{1i} x_{2i} - n\bar{x}_1 \bar{x}_2\right) + \ldots$$

$$\ldots + b_m \left(\sum_{i=1}^{n} x_{1i} x_{mi} - n\bar{x}_1 \bar{x}_m\right) = \sum_{i=1}^{n} x_{1i} y_i - n\bar{x}_1 \bar{y},$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$b_1 \left(\sum_{i=1}^{n} x_{1i} x_{mi} - n\bar{x}_1 \bar{x}_m\right) + b_2 \left(\sum_{i=1}^{n} x_{2i} x_{mi} - n\bar{x}_2 \bar{x}_m\right) +$$

$$+ \ldots + b_m \left(\sum_{i=1}^{n} x_{mi}^2 - n\bar{x}_m \bar{x}_m\right) = \sum_{i=1}^{n} x_{mi} y_i - n\bar{x}_m \bar{y}.$$

Пусть C — матрица коэффициентов при неизвестных параметрах $b_1, b_2, \ldots, b_m$; C^{-1} — матрица, обратная матри-

це C; C_{ij} — элемент, стоящий на пересечении i-й строки и i-го столбца матрицы C^{-1}; A_{yj} — выражение $\sum_{i=1}^{n} x_{ji}y_i - -nx_jy$ $(j=1, 2, ..., m)$. Тогда, используя формулы линейной алгебры, запишем окончательные выражения для параметров:

$$b_j = \sum_{i=1}^{n} C_{ij}^{-1} A_{yi},\ b_0 = \bar{y} - b_1\bar{x}_1 - \ldots - b_m\bar{x}_m. \quad (10.17)$$

Оценкой остаточной дисперсии $\sigma^2_{ост}$ является

$$s^2_{ост} = \left\{\sum_{i=1}^{n} [y_i - y(x_i)]^2\right\}/(n-m-1),$$

где y_i — измеренное значение результативного признака; $y(x_i)$ — значение результативного признака, вычисленное по уравнению регрессии.

Если выборка получена из нормально распределенной генеральной совокупности, то, аналогично изложенному в § 10.4, можно проверить значимость оценок коэффициентов регрессии, только в данном случае статистику $t=|b_j/s_{b_j}|$ вычисляют для каждого j-го коэффициента регрессии $(j=0, 1, 2, ..., m)$:

$$s_{b_0} = s_{ост}\sqrt{1/n + \sum_{m} x_j x_k C_{ij}^{-1}},\ s_{b_j} = s_{ост}\sqrt{C_{jj}^{-1}}$$
$$(i=1, 2, \ldots, n), \quad (10.18)$$

где C_{ij}^{-1} — элемент обратной матрицы, стоящий на пересечении i-й строки и j-го столбца; C_{jj}^{-1} — диагональный элемент обратной матрицы.

При заданном уровне значимости α и числе степеней свободы $k=n-m-1$ по табл. 5 приложений находят критическое значение $t_{\alpha,k}$. Если $t>t_{\alpha k}$, то нулевую гипотезу о равенстве нулю коэффициента регрессии отвергают. Оценку коэффициента считают значимой. Такую проверку производят последовательно для каждого коэффициента регрессии. Если $t<t_{\alpha,k}$, то нет оснований отвергать нулевую гипотезу, оценку коэффициента регрессии считают незначимой.

Для значимых коэффициентов регресссии целесообразно построить доверительные интервалы по формуле (10.10). Для оценки значимости уравнения регрессии

следует проверить нулевую гипотезу о том, что все коэффициенты регрессии (кроме свободного члена) равны нулю: H_0:$\beta^*=0$ (β^* — вектор коэффициентов регрессии). Нулевую гипотезу проверяют, так же как и в § 10.6, с помощью статистики $F=(Q_1/Q_{ост})\,(k_2/k_1)$, где Q_1 — сумма квадратов, характеризующая влияние признаков X; $Q_{ост}$ — остаточная сумма квадратов, характеризующая влияние неучтенных факторов; $k_2=n-m-1$, $k_1=m$. Для уровня значимости α и числа степеней свободы k_1 и k_2 по табл. 4 приложений находят критическое значение $F_{(\alpha,k_1,k_2)}$. Если $F>F_{(\alpha,k_1,k_2)}$, то нулевую гипотезу об одновременном равенстве нулю коэффициентов регрессии отвергают. Уравнение регрессии считают значимым. При $F<F_{(\alpha,k_1,k_2)}$ нет оснований отвергать нулевую гипотезу, уравнение регрессии считают незначимым.

Пример 10.7. Рассчитать уравнение регрессии. Зависимость между результативным признаком Y и признаками X_1 и X_2 имеет вид $M(Y|X=x)=\beta_0+\beta_1x_1+\beta_2x_2$. Оценить значимость уравнения регрессии. Исходные данные приведены в табл. 10.8.

Р е ш е н и е. Кроме исходных данных в табл. 10.8 приведены результаты вычислений, необходимые для составления системы нормальных уравнений. В данном случае нужно оценить параметры β_0, β_1, β_2, а также остаточную дисперсию $\sigma^2_{ост}$. Составим систему нормальных уравнений:

$$\begin{cases} nb_0+b_1\sum\limits_{i=1}^{20}x_{1i}+b_2\sum\limits_{i=1}^{20}x_{2i}=\sum\limits_{i=1}^{20}y_i, \\ b_0\sum\limits_{i=1}^{20}x_{1i}+b_1\sum\limits_{i=1}^{20}x_{1i}^2+b_2\sum\limits_{i=1}^{20}x_{1i}x_{2i}=\sum\limits_{i=1}^{20}x_{1i}y_i, \\ b_0\sum\limits_{i=1}^{20}x_{2i}+b_1\sum\limits_{i=1}^{20}x_{1i}x_{2i}+b_2\sum\limits_{i=1}^{20}x_{2i}^2=\sum\limits_{i=1}^{20}x_{2i}y_i, \end{cases}$$

или

$$\begin{cases} 20b_0+837{,}2b_1+511{,}1b_2=1459{,}2, \\ 837{,}2b_0+36555{,}62b_1+21964{,}13b_2=59745{,}26, \\ 511{,}1b_0+21964{,}13b_1+13830{,}77b_2=36602{,}03. \end{cases}$$

Из первого уравнения системы выразим $b_0=72{,}96-41{,}86\,b_1-25{,}55\,b_2$.

Разделим все члены второго и третьего уравнения системы на коэффициенты при первых членах и подставим полученное выше выражение для b_0 в эти уравнения:

$$\begin{cases} 1{,}8041b_1+0{,}6802b_2=-1{,}5969, \\ 1{,}1142b_1+1{,}5057b_2=-1{,}3458. \end{cases}$$

Таблица 10.8

y	x_1	x_2	yx_1	yx_2	x_1x_2	x_1^2	x_2^2
85,5	25,5	17,3	2180,25	1479,15	441,15	650,25	229,29
81,7	34,1	19,8	2785,97	1617,66	665,18	1162,81	392,04
71,7	37,3	30,1	2674,41	2158,17	1122,73	1391,29	906,01
62,7	44,7	31,9	2802,69	2000,13	1425,93	1998,09	1017,61
66,4	44,6	38,3	2961,44	2543,12	1708,18	1989,16	1466,89
70,6	41,0	26,5	2894,60	1870,90	1086,50	1681,00	702,25
65,0	49,5	36,2	3217,80	2353,0	1791,90	2450,25	1310,44
72,8	45,1	21,0	3283,28	1528,80	947,10	2034,01	441,00
67,6	56,5	29,5	3819,40	1994,20	1666,75	3192,25	870,25
90,0	35,4	29,4	3186,00	2241,00	881,46	1253,16	620,01
60,2	54,9	25,0	3304,98	1505,00	1372,50	3014,01	625,00
74,8	32,8	28,0	2453,44	2094,40	918,40	1075,84	784,00
63,4	55,5	33,9	3518,70	2149,26	1881,45	3080,25	1149,21
74,2	41,5	16,1	3079,30	1194,62	668,15	1722,25	259,21
71,6	41,5	21,0	2971,40	1503,60	871,50	1722,25	441,00
60,0	52,7	28,7	3162,00	1722,00	1512,49	2777,49	823,69
81,1	37,9	20,3	3073,69	1646,33	769,37	1436,41	412,09
71,5	43,9	19,9	3138,83	1422,85	873,61	1927,21	396,01
77,2	35,0	22,6	2702,00	1744,72	791,00	1225,00	510,76
91,2	27,8	20,1	2535,36	1833,12	558,78	772,84	404,01
Σ 1459,2	837,2	511,1	59745,26	36602,03	21964,13	36555,62	13830,77

Найдем b_1 и b_2 методом обратной матрицы. Имеем матрицу коэффициентов и вектор свободных членов

$$C = \begin{bmatrix} 1,8041 & 0,6802 \\ 1,1142 & 1,5057 \end{bmatrix}, \quad Y = \begin{vmatrix} -1,5969 \\ -1,3458 \end{vmatrix}.$$

Находим обратную матрицу коэффициентов:

$$C^{-1} = \begin{vmatrix} 0,7687 & -0,5688 \\ -0,3472 & 0,9211 \end{vmatrix}.$$

Тогда, согласно формуле (10.17),

$$b_1 = 0,7687 \cdot (-1,5969) + 0,3472 \cdot (-1,3458) = 0,7603,$$

$$b_2 = (-0,5688)(-1,5969) + 0,9211(-1,3458) = 0,3313,$$

$$b_0 = 72,96 - (41,86 \cdot 0,7603) + 25,555 \cdot 0,3313 = 113,2457.$$

Уравнение регрессии имеет вид

$$\bar{y}(x) = 113,2457 - 0,7603\, x_1 - 0,3313\, x_2.$$

Оценим остаточную дисперсию. Предварительно вычислим по уравнению регрессии для каждой пары значений x_{i1} и x_{i2} величину $y(x_i)$. Все вычисления сведем в табл. 10.9. Имеем:

$$s^2_{\text{ост}} = \sum_1^{20} [y_i - y(x_i)]^2/(n - m - 1) = 375,211/17 = 22,07,$$

$$s_{\text{ост}} = 4,7.$$

Таблица 10.9

y_i	$y(x_i)$	$y_i-y(x_i)$	$[y_i-y(x_i)]^2$	$(y_i-\bar{y})^2$
85,50	88,13	—2,63	6,917	157,452
81,70	80,77	0,93	0,872	76,213
71,70	74,92	—3,22	10,368	1,588
62,70	68,69	—5,99	35,880	105,468
66,40	66,65	—0,25	0,062	43,034
70,60	73,30	—2,70	7,290	5,570
65,00	63,62	1,38	1,885	63,521
72,80	72,00	0,80	0,640	0,026
67,60	60,53	7,07	49,985	28,837
90,00	78,09	11,91	141,848	290,021
60,20	63,23	—3,03	9,181	162,818
74,80	79,03	—4,23	17,893	3,386
63,40	59,82	3,57	12,766	91,585
74,20	76,37	—2,17	4,709	1,513
71,60	74,75	—3,15	9,922	1,877
60,00	63,67	—3,67	13,469	167,962
81,10	77,71	3,39	11,485	66,260
71,50	73,29	—1,79	3,204	2,161
77,20	79,16	—1,96	3,841	17,893
91,20	85,45	5,75	32,994	332,689
Σ			375,211	1619,483

Оценим значимость уравнения регрессии. Для этого необходимо проверить нулевую гипотезу $H_0: \beta^*=0$. Вычислим статистику $F=(Q_1/Q_{ост})(k_2/k_1)$. Из табл. 10.7 имеем: $Q_{ост}=375{,}211$, $Q=1619{,}478$; отсюда $Q_1=Q-Q_{ост}=1244{,}267$. Число степеней свободы: $k_2=n-m-1=17$, $k_1=m=2$. Далее находим $F=(1244{,}267\times 17)/(375{,}211\cdot 2)=28{,}2$. По табл. 4 приложений для уровня значимости $\alpha=0{,}05$ находим критическое значение $F_{(0{,}05;\ 2;\ 17)}=3{,}60$. $28{,}2>3{,}60$, следовательно, нулевую гипотезу отвергаем, уравнение регрессии считаем значимым.

§ 10.8. Факторный анализ

Основные положения. В последнее время все более широкое распространение находит один из новых разделов многомерного статистического анализа — факторный анализ. Первоначально этот метод разрабатывался для объяснения многообразия корреляций между исходными параметрами. Действительно, результатом корреляционного анализа является матрица коэффициентов кор-

реляций. При малом числе параметров можно произвести визуальный анализ этой матрицы. С ростом числа параметра (10 и более) визуальный анализ не дает положительных результатов. Оказалось, что все многообразие корреляционных связей можно объяснить действием нескольких обобщенных факторов, являющихся функциями исследуемых параметров, причем сами обобщенные факторы при этом могут быть и неизвестны, однако их можно выразить через исследуемые параметры.

Один из основоположников факторного анализа Л. Терстоун приводит такой пример: несколько сотен мальчиков выполняют 20 разнообразных гимнастических упражнений. Каждое упражнение оценивают баллами. Можно рассчитать матрицу корреляций между 20 упражнениями. Это большая матрица размером 20×20. Изучая такую матрицу, трудно уловить закономерность связей между упражнениями. Нельзя ли объяснить скрытую в таблице закономерность действием каких-либо обобщенных факторов, которые в результате эксперимента непосредственно не оценивались? Оказалось, что обо всех коэффициентах корреляции можно судить по трем обобщенным факторам, которые и определяют успех выполнения всех 20 гимнастических упражнений: чувство равновесия, усилие правого плеча, быстрота движения тела.

Дальнейшие разработки факторного анализа показали, что этот метод может быть с успехом применен в задачах группировки и классификации объектов. Факторный анализ позволяет группировать объекты со сходными сочетаниями признаков и группировать признаки с общим характером изменения от объекта к объекту. Действительно, выделенные обобщенные факторы можно использовать как критерии при классификации мальчиков по способностям к отдельным группам гимнастических упражнений.

Методы факторного анализа находят применение в психологии и экономике, социологии и экономической географии. Факторы, выраженные через исходные параметры, как правило, легко интерпретировать как некоторые существенные внутренние характеристики объектов.

Факторный анализ может быть использован и как самостоятельный метод исследования, и вместе с другими методами многомерного анализа, например в соче-

танин с регрессионным анализом. В этом случае для набора зависимых переменных находят обобщенные факторы, которые потом входят в регрессионный анализ в качестве переменных. Такой подход позволяет сократить число переменных в регрессионном анализе, устранить коррелированность переменных, уменьшить влияние ошибок и в случае ортогональности выделенных факторов значительно упростить оценку значимости переменных.

Представление информации в факторном анализе. Для проведения факторного анализа информация должна быть представлена в виде двумерной таблицы чисел размерностью $m \times n$, аналогичной приведенной в § 10.7 (матрица исходных данных). Строки этой матрицы должны соответствовать объектам наблюдений ($i=1, 2, ..., n$) столбцы — признакам ($j=1, 2, ..., m$); таким образом, каждый признак является как бы статистическим рядом, в котором наблюдения варьируют от объекта к объекту. Признаки, характеризующие объект наблюдения, как правило, имеют различную размерность. Чтобы устранить влияние размерности и обеспечить сопоставимость признаков, матрицу исходных данных обычно нормируют, вводя единый масштаб. Самым распространенным видом нормировки является стандартизация. От переменных x_{ji} переходят к переменным $z_{ji}=(x_{ji}-x)/s_j$. В дальнейшем, говоря о матрице исходных переменных, всегда будем иметь в виду стандартизованную матрицу.

Основная модель факторного анализа. Основная модель факторного анализа имеет вид

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \ldots + a_{ip}F_p + d_jU_j, \quad (10.19)$$

где z_j—j-й признак (величина случайная); $F_1, F_2, ..., F_p$ — общие факторы (величины случайные, имеющие нормальный закон распределения); U_j — характерный фактор; $a_{j1}, a_{j2}, ..., a_{jp}$ — факторные нагрузки, характеризующие существенность влияния каждого фактора (параметры модели, подлежащие определению); d_j — нагрузка характерного фактора.

Модель предполагает, что каждый из j признаков, входящих в исследуемый набор и заданных в стандартной форме, может быть представлен в виде линейной комбинации небольшого числа общих факторов F_1, F_2, ..., $F_p (p<m)$ и характерного фактора U_j.

Термин «общий фактор» подчеркивает, что каждый такой фактор имеет существенное значение для анализа всех признаков z_j ($j=1, 2, ..., m$), т. е. F_k — фактор, общий для всех z_j ($k=1, 2, ..., p$).

Термин «характерный фактор» показывает, что он относится только к данному j-му признаку. Это специфика признака, которая не может быть выражена через факторы F_k.

Факторные нагрузки $a_{j1}, a_{j2}, ..., a_{jp}$ характеризуют величину влияния того или иного общего фактора в вариации данного признака. Основная задача факторного анализа — определение факторных нагрузок. Факторная модель относится к классу аппроксимационных. Параметры модели должны быть выбраны так, чтобы наилучшим образом аппроксимировать корреляции между наблюдаемыми признаками.

Для j-го признака и i-го объекта модель (10.19) можно записать в виде

$$z_{ji} = \sum_{k=1}^{p} a_{jk} F_{ki} + d_j U_{ji} \quad (i=1, 2, ..., n,$$

$$j = 1, 2, \ldots, m), \qquad (10.20)$$

где F_{ki} — значение k-го фактора для i-го объекта.

Дисперсию признака s_j^2 можно разложить на составляющие: часть, обусловленную действием общих факторов, — общность h_j^2 и часть, обусловленную действием j-го характерного фактора, — *характерность* d_j^2. Все переменные представлены в стандартизованном виде, поэтому дисперсия j-го признака $s_j^2 = 1$. Дисперсия признака может быть выражена через факторы и в конечном счете через факторные нагрузки.

Если общие и характерные факторы не коррелируют между собой, то дисперсию j-го признака можно представить в виде

$$s_j^2 = 1 = \underbrace{a_{j1}^2 + a_{j2}^2 + \ldots + a_{jk}^2 + \ldots + a_{jp}^2}_{h_j^2} + d_j^2,$$

где a_{jk}^2 — доля дисперсии признака z_j, приходящаяся на k-й фактор.

Полный вклад k-го фактора в суммарную дисперсию признаков

$$V_k = \sum_{j=1}^{m} a_{jk}^2 \ (k = 1, 2, \ldots, p).$$

Вклад общих факторов в суммарную дисперсию $V = \sum_{k=1}^{p} V_k$.

Факторное отображение. Используя модель (10.19), запишем выражения для каждого из параметров:

$$\begin{cases} z_1 = a_{11} F_1 + a_{12} F_2 + \ldots + a_{1p} F_p + d_1 U_1, \\ z_2 = a_{21} F_1 + a_{22} F_2 + \ldots + a_{2p} F_p + d_2 U_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ z_m = a_{m1} F_1 + a_{m2} F_2 + \ldots + a_{mp} F_p + d_m U_m. \end{cases} \tag{10.21}$$

Коэффициенты системы (10.21) — факторные нагрузки — можно представить в виде матрицы, каждая строка которой соответствует параметру, а столбец — фактору.

Факторный анализ позволяет получить не только матрицу отображений, но и коэффициенты корреляции между параметрами и факторами, что является важной характеристикой качества факторной модели. Таблица таких коэффициентов корреляции называется *факторной структурой* или просто *структурой*.

Коэффициенты отображения можно выразить через выборочные парные коэффициенты корреляции. На этом основаны методы вычисления факторного отображения.

Рассмотрим связь между элементами структуры и коэффициентами отображения. Для этого, учитывая выражение (10.19) и определение выборочного коэффициента корреляции, умножим уравнения системы (10.21) на соответствующие факторы, произведем суммирование по всем n наблюдениям и, разделив на n, получим следующую систему уравнений:

$$\begin{cases} r_{z_j F_1} = a_{j1} + a_{j2} r_{F_1 F_2} + \ldots + a_{jp} r_{F_1 F_p}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ r_{z_j F_k} = a_{j1} r_{F_k F_1} + a_{j2} r_{F_k F_2} + \ldots + a_{jp} r_{F_k F_p}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ r_{z_j F_p} = a_{j1} r_{F_p F_1} + a_{j2} r_{F_p F_2} + \ldots + a_{jp}, \\ r_{z_j U} = d_j, \end{cases} \tag{10.22}$$

где $r_{z_j F_k}$ — выборочный коэффициент корреляции между j-м параметром и k-м фактором; $r_{F_k F_p}$ — коэффициент корреляции между k-м и p-м факторами.

Если предположить, что общие факторы между собой не коррелированы, то уравнения (10.22) можно записать в виде $r_{z_j F_k} = a_{jk}$ ($j=1, 2, ..., m$; $k=1, 2, ..., p$), т. е. коэффициенты отображения равны элементам структуры.

Введем понятие *остаточного коэффициента корреляции* и *остаточной корреляционной матрицы*. Исходной информацией для построения факторной модели (10.19) служит матрица выборочных парных коэффициентов корреляции. Используя построенную факторную модель, можно снова вычислить коэффициенты корреляции между признаками и сравнить их с исходными коэффициентами корреляции. Разница между ними и есть остаточный коэффициент корреляции.

В случае независимости факторов имеют место совсем простые выражения для вычисляемых коэффициентов корреляции между параметрами: для их вычисления достаточно взять сумму произведений коэффициентов отображения, соответствующих наблюдавшимся признакам:

$$r'_{jk} = a_{j1} a_{k1} + a_{j2} a_{k2} + ... + a_{jp} a_{kp} \; (j \neq k, j, k = 1, 2, ..., p),$$

где r'_{jk} — вычисленный по отображению коэффициент корреляции между j-м и k-м признаком. Остаточный коэффициент корреляции

$$\bar{r}_{jk} = r_{jk} - r'_{jk}.$$

Матрица остаточных коэффициентов корреляции называется *остаточной матрицей* или *матрицей остатков*

$$\bar{R} = R - R',$$

где $\bar{R}$ — матрица остатков; R — матрица выборочных парных коэффициентов корреляции, или полная матрица; R' — матрица вычисленных по отображению коэффициентов корреляции.

Результаты факторного анализа удобно представить в виде табл. 10.10.

Здесь суммы квадратов нагрузок по строкам — общности параметров, а суммы квадратов нагрузок по столбцам — вклады факторов в суммарную дисперсию параметров. Имеет место соотношение

Таблица 10.10

	Факторные нагрузки	Общности
A	$\begin{matrix} a_{11}\ a_{21} \ldots a_{k1} \ldots a_{p1} \\ \ldots\ldots\ldots\ldots \\ a_{1m}\ a_{2m} \ldots a_{km} \ldots a_{pm} \end{matrix}$	$\begin{matrix} h_1^2 \\ \vdots \\ h_m^2 \end{matrix}$
Вклады факторов	$V_1\ V_2 \ldots V_k \ldots V_p$	V

$$\sum_{k=1}^{p} V_k = \sum_{j=1}^{m} h_j^2 = \sum_{ik} a_{ik}^2 = V.$$

Определение факторных нагрузок. Матрицу факторных нагрузок можно получить различными способами. В настоящее время наибольшее распространение получил *метод главных факторов*. Этот метод основан на принципе последовательных приближений и позволяет достичь любой точности. Метод главных факторов предполагает использование ЭВМ. Существуют хорошие алгоритмы и программы, реализующие все вычислительные процедуры.

Введем понятие редуцированной корреляционной матрицы или просто редуцированной матрицы. *Редуцированной* называется матрица выборочных коэффициентов корреляции $\widetilde{R}$, у которой на главной диагонали стоят значения общностей h_k^2 $(k=1, 2, \ldots, p)$:

$$\widetilde{R} = \begin{bmatrix} h_1^2 & r_{21} & r_{31} & \ldots & r_{m1} \\ r_{12} & h_2^2 & r_{32} & \ldots & r_{m2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{1m} & r_{2m} & \ldots & & h_m^2 \end{bmatrix}.$$

Редуцированная и полная матрицы связаны соотношением

$$\widetilde{R} = R - D, \qquad (10.23)$$

где D — матрица характерностей.

Общности, как правило, неизвестны, и нахождение их в факторном анализе представляет серьезную проблему. Вначале определяют (хотя бы приближенно) число общих факторов, совокупность которых может с достаточной точностью аппроксимировать все взаимосвязи выборочной корреляционной матрицы. Доказано, что число общих факторов (общностей) равно рангу редуцированной матрицы, а при известном ранге можно по выборочной корреляционной матрице найти оценки общностей. Числа общих факторов можно определить априори, исходя из физической природы эксперимента. Затем рассчитывают матрицу факторных нагрузок. Такая матрица, рассчитанная методом главных факторов, обладает одним интересным свойством: сумма произведений каждой пары ее столбцов равна нулю, т. е. факторы попарно ортогональны.

Сама процедура нахождения факторных нагрузок, т. е. матрицы A, состоит из нескольких шагов и заключается в следующем: на первом шаге ищут коэффициенты факторных нагрузок при первом факторе так, чтобы сумма вкладов данного фактора в суммарную общность была максимальной:

$$V_1 = a_{11}^2 + a_{21}^2 + \ldots + a_{p1}^2. \qquad (10.24)$$

Максимум V_1 должен быть найден при условии

$$r_{j\xi} = \sum_{k=1}^{p} a_{jk} a_{\xi k} \qquad (j, \xi = 1, 2, \ldots, m), \qquad (10.25)$$

где $r_{j\xi} = r_{\xi j}$; r_{jj} — общность h_j^2 параметра z_j.

Затем рассчитывают матрицу коэффициентов корреляции с учетом только первого фактора $R_1' = a_1 a_1$. Имея эту матрицу, получают первую матрицу остатков: $R_1 = \tilde{R} - R_1'$.

На втором шаге определяют коэффициенты нагрузок при втором факторе так, чтобы сумма вкладов второго фактора в остаточную общность (т. е. полную общность без учёта той части, которая приходится на долю первого фактора) была максимальной. Сумма квадратов нагрузок при втором факторе

$$V_2 = a_{12}^2 + a_{22}^2 + \ldots + a_{p2}^2. \qquad (10.26)$$

Таблица 10.10

	Факторные нагрузки	Общности
A	$a_{11}\ a_{21} \ldots a_{k1} \ldots a_{p1}$ $\ldots\ldots\ldots\ldots$ $a_{1m}\ a_{2m} \ldots a_{km} \ldots a_{pm}$	h_1^2 $\vdots$ h_m^2
Вклады факторов	$V_1\ V_2 \ldots V_k \ldots V_p$	V

$$\sum_{k=1}^{p} V_k = \sum_{j=1}^{m} h_j^2 = \sum_{ik} a_{ik}^2 = V.$$

Определение факторных нагрузок. Матрицу факторных нагрузок можно получить различными способами. В настоящее время наибольшее распространение получил *метод главных факторов*. Этот метод основан на принципе последовательных приближений и позволяет достичь любой точности. Метод главных факторов предполагает использование ЭВМ. Существуют хорошие алгоритмы и программы, реализующие все вычислительные процедуры.

Введем понятие редуцированной корреляционной матрицы или просто редуцированной матрицы. *Редуцированной* называется матрица выборочных коэффициентов корреляции $\bar{R}$, у которой на главной диагонали стоят значения общностей h_k^2 $(k=1, 2, \ldots, p)$:

$$\bar{R} = \begin{bmatrix} h_1^2 & r_{21} & r_{31} & \ldots & r_{m1} \\ r_{12} & h_2^2 & r_{32} & \ldots & r_{m2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{1m} & r_{2m} & \ldots & & h_m^2 \end{bmatrix}.$$

Редуцированная и полная матрицы связаны соотношением

$$\bar{R} = R - D, \qquad (10.23)$$

где D — матрица характерностей.

Общности, как правило, неизвестны, и нахождение их в факторном анализе представляет серьезную проблему. Вначале определяют (хотя бы приближенно) число общих факторов, совокупность которых может с достаточной точностью аппроксимировать все взаимосвязи выборочной корреляционной матрицы. Доказано, что число общих факторов (общностей) равно рангу редуцированной матрицы, а при известном ранге можно по выборочной корреляционной матрице найти оценки общностей. Числа общих факторов можно определить априори, исходя из физической природы эксперимента. Затем рассчитывают матрицу факторных нагрузок. Такая матрица, рассчитанная методом главных факторов, обладает одним интересным свойством: сумма произведений каждой пары ее столбцов равна нулю, т. е. факторы попарно ортогональны.

Сама процедура нахождения факторных нагрузок, т. е. матрицы A, состоит из нескольких шагов и заключается в следующем: на первом шаге ищут коэффициенты факторных нагрузок при первом факторе так, чтобы сумма вкладов данного фактора в суммарную общность была максимальной:

$$V_1 = a_{11}^2 + a_{21}^2 + \ldots + a_{p1}^2. \qquad (10.24)$$

Максимум V_1 должен быть найден при условии

$$r_{j\xi} = \sum_{k=1}^{p} a_{jk} a_{\xi k} \qquad (j, \xi = 1, 2, \ldots, m), \qquad (10.25)$$

где $r_{j\xi} = r_{\xi j}$; r_{jj} — общность h_j^2 параметра z_j.

Затем рассчитывают матрицу коэффициентов корреляции с учетом только первого фактора $R_1' = a_1 a_1'$. Имея эту матрицу, получают первую матрицу остатков: $R_1 = \tilde{R} - R_1'$.

На втором шаге определяют коэффициенты нагрузок при втором факторе так, чтобы сумма вкладов второго фактора в остаточную общность (т. е. полную общность без учёта той части, которая приходится на долю первого фактора) была максимальной. Сумма квадратов нагрузок при втором факторе

$$V_2 = a_{12}^2 + a_{22}^2 + \ldots + a_{p2}^2. \qquad (10.26)$$

Максимум V_2 находят из условия

$$\bar{r}^1_{j\xi} = \sum_{k=1}^{p} a^1_{jk} a^1_{\xi k} \quad (j, \xi = 1, 2, \ldots, m), \tag{10.27}$$

где $\bar{r}^1_{j\xi}$ — коэффициент корреляции из первой матрицы остатков; $a^1_{jk}, a^1_{\xi k}$ — факторные нагрузки с учетом второго фактора. Затем рассчитывают матрицу коэффициентов корреляций с учетом второго фактора и вычисляют вторую матрицу остатков: $\bar{R}_2 = \bar{R}_1 - R_2$.

Факторный анализ учитывает суммарную общность. Исходная суммарная общность $H = \sum_{j=1}^{m} h_j^2$. Итерационный процесс выделения факторов заканчивают, когда учтенная выделенными факторами суммарная общность отличается от исходной суммарной общности меньше чем на ε (ε — наперед заданное малое число).

Адекватность факторной модели оценивается по матрице остатков (если величины ее коэффициентов малы, то модель считают адекватной).

Такова последовательность шагов для нахождения факторных нагрузок. Для нахождения максимума функции (10.24) при условии (10.25) используют метод множителей Лагранжа, который приводит к системе m уравнений относительно m неизвестных a_{jl}.

Метод главных компонент. Разновидностью метода главных факторов является *метод главных компонент* или *компонентный анализ*, который реализует модель вида

$$z_j = a_{j1} F_1 + a_{j2} F_2 + \ldots + a_{jm} F_m, \tag{10.28}$$

где m — количество параметров (признаков).

Каждый из наблюдаемых параметров линейно зависит от m не коррелированных между собой новых компонент (факторов) $F_1, F_2, \ldots, F_m$. По сравнению с моделью факторного анализа (10.19) в модели (10.28) отсутствует характерный фактор, т. е. считается, что вся вариация параметра может быть объяснена только действием общих или главных факторов. В случае компонентного анализа исходной является матрица коэффициентов корреляции, где на главной диагонали стоят единицы. Результатом компонентного анализа, так же как и факторного, является матрица факторных нагру-

зок. Поиск факторного решения — это ортогональное преобразование матрицы исходных переменных, в результате которого каждый параметр может быть представлен линейной комбинацией найденных m факторов, которые называют *главными компонентами*. Главные компоненты легко выражаются через наблюденные параметры.

Если для дальнейшего анализа оставить все найденные m компонент, то тем самым будет использована вся информация, заложенная в корреляционной матрице. Однако это неудобно и нецелесообразно. На практике обычно оставляют небольшое число компонент, причем количество их определяется долей суммарной дисперсии, учитываемой этими компонентами. Существуют различные критерии для оценки числа оставляемых компонент; чаще всего используют следующий простой критерий: оставляют столько компонент, чтобы суммарная дисперсия, учитываемая ими, составляла заранее установленное число процентов. Первая из компонент должна учитывать максимум суммарной дисперсии параметров; вторая — не коррелировать с первой и учитывать максимум оставшейся дисперсии и так до тех пор, пока вся дисперсия не будет учтена. Сумма учтенных всеми компонентами дисперсий равна сумме дисперсий исходных параметров. Математический аппарат компонентного анализа полностью совпадает с аппаратом метода главных факторов. Отличие только в исходной матрице корреляций.

Компонента (или фактор) через исходные переменные выражается следующим образом:

$$F_{ik} = (a_{j1} z_{j1} + a_{j2} z_{j2} + \ldots + a_{jm} z_{jm})/\lambda_k$$
$$(i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m, \quad k = 1, 2, \ldots, p), \qquad (10.29)$$

где a_{jm} — элементы факторного решения: z_{jm} — исходные переменные; λ_k — k-е собственное значение; p — количество оставленных главных компонент.

Для иллюстрации возможностей факторного анализа покажем, как, используя метод главных компонент, можно сократить размерность пространства независимых переменных, перейдя от взаимно коррелированных параметров к независимым факторам, число которых $p < m$.

Пример 10.8. Исследуется зависимость себестоимости покрышек для автомобилей от технико-экономических параметров. Найдем вид этой зависимости.

Таблица 10.11

z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8
1	0,9806	0,9819	0,9543	0,9456	0,8703	—0,3848	—0,1502
		0,9936	0,9771	0,9414	0,8677	—0,4130	—0,1460
		1	0,9566	0,9620	0,9000	—0,4344	—0,1130
			1	0,8840	0,7921	—0,3711	—0,2147
				1	0,9712	—0,4452	—0,0854
					1	—0,5133	—0,0223
						1	—0,0437
							1

Взяты следующие параметры: x_1 — масса покрышки; x_2 — наружный диаметр; x_3 — ширина профили; x_4 — статический радиус; x_5 — максимально допустимая нагрузка; x_6 — норма слойности; x_7 — объем выпуска; x_8 — гарантийная наработка. Имеется 60 объектов.

Матрица технико-экономических параметров предварительно стандартизируется. Вектор себестоимости обозначим Y. Требуемую зависимость можно получить непосредственно, найдя уравнение регрессии Y на X. Однако анализ корреляционной матрицы (табл. 10.11) показывает сильную взаимную коррелированность между параметрами. Факторный анализ позволяет перейти от взаимозависимых параметров к независимым факторам, число которых $p<m$. По матрице корреляций, вычислены собственные векторы и собственные значения и в итоге найдена матрица факторных нагрузок (табл. 10.12). Подсчитан также процент суммарной дисперсии, учитываемой каждой компонентой. Первые четыре компоненты учитывают более 99% суммарной дисперсии параметров, поэтому в дальнейшем следует рассматривать только их. Таким образом, всю вариацию параметров можно объяснить действием четырех обобщенных факторов, которые не были измерены и включены в исходную информацию.

Компоненты достаточно просто интерпретируются. Рассмотрим матрицу факторных нагрузок (табл. 10.12). Первый столбец соответствует первой компоненте. Здесь вклады почти всех параметров в компоненту достаточно высоки и примерно одинаковы. Параметры — это технические характеристики изделия, следовательно, первую компоненту можно считать обобщенным фактором технических характеристик изделия. Второй столбец соответствует второй компоненте. Здесь наибольший вклад дает компонента x_8— объем выпуска. Следовательно, вторую компоненту можно считать обобщенным производственным фактором. Третью компоненту можно интерпретировать как обобщенный фактор эксплуатационных характеристик. Четвертая компонента — более узкая характеристика технических параметров изделия, отражающая его геометрию. Ее можно считать фактором геометрических характеристик покрышки.

Теперь по формуле (10.19) для каждого объекта вычислим значения всех четырех компонент. В результате получим новую матрицу F размерностью $p \times n$. Найдем теперь зависимость между Y и обобщенными факторами F. Для этого применим обычный регрес-

Таблица 10.12

Матрица факторных нагрузок

	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	Дисперсии	Дисперсия, %
z_1	0,9714	—0,1293	0,1068	—0,0841	—0,1287	0,0238	0,0626	—0,0013	6,56	72,01
z_2	0,9766	—0,1295	0,0790	—0,1345	0,0196	0,0315	—0,0461	—0,0278	1,34	14,89
z_3	0,9876	—0,0840	0,0653	—0,0730	—0,0048	0,0612	—0,056	0,0083	0,77	8,60
z_4	0,9338	—0,2200	0,0907	—0,2382	0,0938	—0,0624	0,0428	0,0083	0,25	2,74
z_5	0,9827	—0,0118	0,0449	0,1693	—0,0066	—0,0202	—0,0224	0,0482	0,03	0,36
z_6	0,9435	0,0821	—0,0517	0,2999	0,0714	0,0538	0,0465	—0,0165	0,02	0,27
z_7	—0,5068	—0,2018	0,8344	0,0780	0,0110	0,0062	0,0008	—0,0013	0,01	0,17
z_8	—0,0556	0,9717	0,1625	—0,1524	0,0220	0,0458	0,0159	0,0125	0,04	0,05
										$\Sigma=99,98$

сионный анализ. Уравнение регрессии имеет вид

$$Y = 94{,}92 + 34{,}66F_1 + 0{,}74F_2 + 0{,}76F_3 - 2{,}46F_4.$$

Все переменные были приведены к одному масштабу, поэтому коэффициенты регрессии позволяют оценить истинный вклад каждого фактора в себестоимость. Так, из уравнения регрессии следует, что максимальный вклад вносит первый обобщенный фактор — фактор технических характеристик изделия. Вклад остальных факторов гораздо меньше.

Для данного уравнения вычислены совокупный коэффициент корреляции $R=0{,}974$ и статистика $F_{выч}=s^2/s^2_{ост}$ для оценки по критерию Фишера. По табл. 5 приложений для $\alpha=0{,}05$ и $k_1=4$, $k_2=60-4-1=55$ находим $F_{(0{,}05;\ 4;\ 55)}=2{,}52$. Оценка точности аппроксимации данным уравнением регрессии показывает, что связь между зависимой переменной и четырьмя независимыми факторами очень тесная ($R=0{,}974$). Статистика $F_{выч}=17{,}93 \gg F_{табл}$, значит, аппроксимация данным уравнением регрессии хорошая.

Следует особо остановиться на интерпретации результатов, т. е. на смысловой стороне факторного анализа. Собственно факторный анализ состоит из двух важных этапов; аппроксимации корреляционной матрицы и интерпретации результатов. Аппроксимировать корреляционную матрицу, т. е. объяснить корреляцию между параметрами действием каких-либо общих для них факторов, и выделить сильно коррелирующие группы параметров достаточно просто: из корреляционной матрицы одним из методов факторного анализа непосредственно получают матрицу нагрузок — факторное решение, которое называют *прямым факторным решением*. Однако часто это решение не удовлетворяет исследователей. Они хотят интерпретировать фактор как скрытый, но существенный параметр, поведение которого определяет поведение некоторой своей группы наблюдаемых параметров, в то время как поведение других параметров определяется поведением других факторов. Для этого у каждого параметра должна быть наибольшая по модулю факторная нагрузка с одним общим фактором. Прямое решение следует преобразовать, что равносильно повороту осей общих факторов. Такие преобразования называют *вращениями*; в итоге получают *косвенное факторное решение*, которое и является результатом факторного анализа.

Глава 11

Планирование эксперимента

§ 11.1. Общая идея планирования эксперимента

Планирование эксперимента — новый раздел математической статистики.

Оценка влияния одного и двух факторов на исследуемую величину рассмотрена в гл. 8. Так, например, при исследовании влияния двух факторов (см. табл. 8.5) количество опытов полного факторного эксперимента составляет rv, где r — число уровней первого фактора, v — второго. При $r=v=2$ количество опытов равно $2\cdot 2=2^2=4$. Если число уровней каждого из факторов одинаково, то количество опытов, которые надо провести по схемам полного факторного эксперимента при исследовании влияния k факторов, можно вычислить по формуле $N=v^k$, где v — число уровней каждого из факторов. На практике бывает необходимо исследовать влияние на рассматриваемую величину, например, десяти факторов, каждый из которых имеет четыре уровня. В этом случае $N=4^{10}=1\,048\,576$ опытов.

Вследствие непомерно большого количества опытов в подобных задачах использовать методы дисперсионного анализа, рассмотренные в гл. 8, практически невозможно. Такого рода задачи явились одной из основных причин возникновения теории планирования эксперимента, которая позволяет ответить на вопрос: сколько и какие опыты следует включить в эксперимент. Родоначальником этого направления является английский ученый Р. А. Фишер. В современной форме планирование эксперимента начало развиваться после выхода в свет работ Г. Е. Бокса, К. В. Уилсона, В. В. Налимова.

Планирование эксперимента начинают с выбора объекта исследования, который изучается с определенной целью (ради отыскания оптимальных условий протекания химических, физических, металлургических и других процессов). Цель исследования называют *целевой функцией*, или *параметром оптимизации*, или *критерием оптимизации*. Способы воздействия на объект ис-

следования, как и в дисперсионном анализе, называют факторами.

Для того чтобы прогнозировать значение целевой функции, необходимо параметр оптимизации связать с факторами некоторой функциональной зависимостью. Эту зависимость, имеющую вид $Y=f(x_1, x_2, ..., x_l)$, называют функцией, или *поверхностью отклика*, или *моделью объекта исследования*.

Компенсацией за меньшее количество опытов по сравнению с полным факторным экспериментом служат ограничения, принимаемые исследователем априори, т. е. до проведения опыта. В качестве ограничений принимают существование единственного оптимума и представление функции отклика в виде полинома заданного порядка, параметры которого оценивают по опытным данным с помощью регрессионного анализа. Если же в действительности модель не удовлетворяет наложенным ограничениям, то оптимум функции отклика можно и не найти.

Рассмотрим теперь, как принятые допущения способствуют уменьшению количества опытов. Пусть, например, известно значение параметра оптимизации в нескольких соседних точках. В силу непрерывности функции отклика можно прогнозировать значения параметра оптимизации в окрестностях соседних точек. Следовательно, можно обнаружить точки, для которых ожидается увеличение (или уменьшение, если отыскивается минимум) параметра оптимизации. В силу единственности оптимума следующий эксперимент целесообразно поставить в точках, в которых обнаружено эффективное изменение параметра оптимизации, пренебрегая всеми остальными. В результате такого пошагового продвижения может быть достигнут оптимум параметра оптимизации.

Направление наибольшей скорости возрастания функции отклика называют *направлением градиента*. Если нет особых указаний о виде целевой функции, то в начале эксперимента всегда используют линейную модель, так как она определяется минимально возможным числом коэффициентов при данном числе факторов. Двигаясь по градиенту, строят линейные модели до тех пор, пока они дают эффективное изменение параметра оптимизации. Если улучшение параметра оптимизации с помощью линейной модели больше не наблюдается, то

обнаружена область, близкая к оптимуму, или «почти стационарная». В этом случае либо исследование прекращают, либо исследуют полиномы более высоких степеней.

§ 11.2. Полный факторный эксперимент

После выбора объекта исследования, формулировки целевой функции и описания факторов перед экспериментатором встает вопрос: при каких сочетаниях факторов проводить первые опыты? При выборе экспериментальной области необходимо использовать априорную информацию об исследуемом процессе. В качестве отправной обычно выбирают точку, соответствующую наилучшим сочетаниям факторов, т. е. такую, при которой значение целевой функции максимально (минимально) по сравнению с другими известными сочетаниями факторов.

Точку начала эксперимента называют *нулевым* или *основным уровнем*. Если априорная информация отсутствует, то выбор нулевого уровня произволен, однако координаты точки начала опыта должны лежать внутри области определения факторов, на некотором расстоянии от границы.

Далее переходят к выбору интервалов варьирования по каждому из факторов. Под интервалом варьирования в данном случае понимают число (свое для каждого фактора), прибавляя которое к нулевому уровню, получают верхний, а вычитая — нижний уровни фактора. На первом этапе планирования эксперимента (при получении линейной модели) факторы всегда варьируют лишь на двух уровнях. Интервал варьирования не может быть меньше ошибки, с которой экспериментатор фиксирует уровень фактора, иначе верхний и нижний уровни окажутся неразличимыми. С другой стороны, нижний и верхний интервалы должны, как и нулевой уровень, лежать внутри области определения факторов. Выбор нулевого уровня и интервалов варьирования — задача трудная, так как она связана с неформализованным этапом планирования эксперимента.

Пусть x_{j0} — нулевой уровень; h_j — интервал варьирования; x_j — значение фактора; j — номер фактора. Для простоты записи и обработки экспериментальных данных перейдем к новой безразмерной системе координат с

началом в центре исследуемой области. В новой системе координат значение j-го фактора обозначим X_j. Значение X_j связано с x_j следующей формулой:

$$X_j = (x_j - x_{j0})/h_j. \qquad (11.1)$$

Используя эту формулу, можно показать, что в новой системе координат x_{j0} принимает значение — 0, верхний уровень $x_{j0}+h_j=x_{jв}$ — значение $+1$, а нижний уровень $x_{j0}-h_j=x_{jн}$ — значение -1.

Пример 11.1. Пусть процесс определяется двумя факторами. Основной уровень и интервалы варьирования выбраны так, как указано в табл. 11.1.

Таблица 11.1

	x_1	x_2
Основной уровень	1,2	3
Интервал варьирования	1	2

В результате опыта получена точка A с кординатами $x_1=2$; $x_2=4$. Необходимо вычислить по каждому из факторов верхний и нижний уровень, кодированные значения (т. е. в новой системе координат) основного, верхнего и нижнего уровней, а также точки A.

Решение. Выбор нулевого уровня и интервала варьирования однозначно определяет верхний и нижний уровни фактора. Имеем:

верхний уровень $x_{10}+h_1=1,2+1=2,2$; $x_{20}+h_2=3+2=5$;
нижний уровень $x_{10}-h_1=1,2-1=0,2$; $x_{20}-h_2=3-2=1$;
кодированные значения таковы:
основного уровня $X_{10}=(1,2-1,2)/1=0$; $X_{20}=(3-3)/2=0$;
верхнего уровня $X_{1в}=(2,2-1,2)/1=+1$; $X_{2в}=(5-3)/2=+1$;
нижнего уровня $X_{1н}=(0,2-1,2)/1=-1$; $X_{2н}=(1-3)/2=-1$;
точки $AX_1=(2-1,2)/1=0,8$; $X_2=(4-3)/2=0,5$.

Эксперимент, в котором реализуются все возможные сочетания уровней факторов, называют *полным* факторным экспериментом. Методы обработки информации полного факторного эксперимента исследуют в дисперсионном анализе.

Если число факторов известно, то при варьировании факторов на двух уровнях количество опытов можно вычислить по формуле

$$N = 2^k, \qquad (11.2)$$

где N — количество опытов, k — количество факторов.

Составим матрицу планирования эксперимента для полного факторного эксперимента 2^2 (табл. 11.2).

Таблица 11.2

№ опыта	X_1	X_2	Y
1	−1	−1	y_1
2	+1	−1	y_2
3	−1	+1	y_3
4	+1	+1	y_4

В матрице планирования указывают все возможные сочетания нижних и верхних уровней по каждому из факторов модели, в последнем столбце записывают значения выходного параметра, соответствующие определенным сочетаниям факторов. Заметим, что иногда в матрице планирования единицы опускают, тогда таблица 11.2 принимает вид табл. 11.3.

Таблица 11.3

№ опыта	X_1	X_2	Y
1	−	−	y_1
2	+	−	y_2
3	−	+	y_3
4	+	+	y_4

Очевидно, что для двух факторов все возможные комбинации уровней легко найти перебором, однако с ростом числа факторов возникает необходимость в другом методе построения матриц планирования. Рассмотрим один из них. При добавлении нового фактора каждая комбинация уровней исходного плана встречается дважды: в сочетании с нижним и верхним уровнями нового фактора.

Рассмотрим этот прием при переходе от эксперимента 2^2 к эксперименту 2^3. Запишем матрицу 2^2 дважды. В столбце X_3 четыре раза запишем знак плюс, а затем ниже — четыре раза — минус (табл. 11.4).

Этим способом можно получить матрицу любой размерности.

Рассмотрим теперь общие свойства матрицы планирования. Как будет показано ниже, эти свойства позволят быстро и просто рассчитывать целевую функцию.

Таблица 11.4

№ опыта	X_1	X_2	X_3	Y
1	−1	−1	+1	y_1
2	+1	−1	+1	y_2
3	−1	+1	+1	y_3
4	+1	+1	+1	y_4
5	−1	−1	−1	y_5
6	+1	−1	−1	y_6
7	−1	+1	−1	y_7
8	+1	+1	−1	y_8

1°. *Симметричность относительно нулевого уровня, т. е. алгебраическая сумма элементов столбца каждого фактора, равна нулю.*

2°. *Сумма квадратов элементов столбца каждого из факторов равна числу опытов (свойство нормировки).*

3°. *Произведение любых двух различных векторстолбцов факторов равно нулю.* При этом каждый столбец в матрице планирования рассматривается как вектор-столбец, а строка — как вектор-строка (*свойство ортогональности*).

4°. *Дисперсии предсказанных значений параметра оптимизации одинаковы на равных расстояниях от нулевого уровня (свойство ротатабельности матрицы планирования).*

После того как по выбранной матрице планирования эксперимент проведен, обычно переходят к оценке параметров целевой функции. Если, например, исследуют два фактора, то функция отклика может иметь вид

$$X = b_0 + b_1 X_1 + b_2 X_2. \qquad (11.3)$$

При условии, что факторы варьируют на двух уровнях, по матрице табл. 11.5 получают числовые значения b_0, b_1, b_2.

Эту матрицу часто называют *расширенной информационной матрицей*, так как по сравнению с матрицей табл. 11.2 в нее введен столбец X_0, состоящий из одних единиц и получивший название *фиктивного*. В § 8.3 сказано, что при проведении полного факторного эксперимента можно количественно оценить не только силу

Таблица 11.5

№ опыта	X_0	X_1	X_2	Y
1	+1	+1	+1	y_1
2	+1	−1	+1	y_2
3	+1	−1	−1	y_3
4	+1	+1	−1	y_4

Таблица 11.6

№ опыта	X_0	X_1	X_2	X_1X_2	Y
1	1	+1	+1	+1	y_1
2	1	−1	+1	−1	y_2
3	1	−1	−1	+1	y_3
4	1	+1	−1	−1	y_4

влияния факторов на параметр оптимизации, но и эффекты взаимодействия, которые часто влияют на целевую функцию (так, например, добавление одних веществ может стимулировать влияние других на объект исследования). Следовательно, при анализе двух факторов полный факторный эксперимент позволяет количественно оценить параметры следующей модели: $Y=b_0+b_1X_1+b_2X_2+b_{12}X_1X_2$, причем для получения значений b_0, b_1, b_{21}, b_{12} необходимо воспользоваться расширенной информационной матрицей в виде табл. 11.6.

Элементы столбца X_1X_2 получают, построчно перемножая соответствующие элементы столбцов X_1 и X_2. Матрица, состоящая из столбцов X_1, X_2, X_1X_2, взятых из табл. 11.6, сохраняет все свойства матрицы эксперимента.

В начале исследования почти всегда используют линейную модель, при этом количество опытов полного факторного эксперимента находят по формуле (11.2). Числовые соотношения между количеством факторов, количеством параметров линейной модели и числом опытов полного факторного эксперимента приведены в табл. 11.7.

Как видно из табл. 11.7, разность между числом опытов и количеством параметров линейной модели с увеличением числа факторов становится непомерно большой. В следующем параграфе рассмотрим прием, поз-

Таблица 11.7

Количество факторов	Количество параметров линейной модели	Число опытов полного факторного эксперимента	Разность между числом опытов и количеством параметров
2	3	4	1
3	4	8	4
4	5	16	11
5	6	32	26
6	7	64	57
7	8	128	120
8	9	256	247
9	10	512	502
10	11	1 024	1 013
11	12	2 048	2 026
12	13	4 096	4 083
13	14	8 192	8 178
14	15	16 384	16 369
15	16	32 768	32 752

воляющий исследовать линейную модель, используя меньшее количество опытов по сравнению с числом опытов полного факторного эксперимента.

§ 11.3. Дробный факторный эксперимент

Пусть, например, для описания объекта исследования требуется рассчитать коэффициенты b_0, b_1, b_2, b_3 уравнения

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3. \qquad (11.4)$$

Для определения числовых значений четырех параметров следует иметь четыре уравнения, неизвестными в которых являются параметры рассматриваемой функции, поэтому как минимум надо провести четыре опыта.

Уравнения можно записать так:

$$\begin{array}{ll} \text{1-й опыт} & y_1 = b_0 + b_1 x_{11} + b_2 x_{21} + b_3 x_{31}, \\ \text{2-й} \quad » & y_2 = b_0 + b_1 x_{12} + b_2 x_{22} + b_3 x_{32}, \\ \text{3-й} \quad » & y_3 = b_0 + b_1 x_{13} + b_2 x_{23} + b_3 x_{33}, \\ \text{4-й} \quad » & y_4 = b_0 + b_1 x_{14} + b_2 x_{24} + b_3 x_{34}, \end{array}$$

где y_i — значение параметра оптимизации в i-м опыте; x_{ji} — значение j-го фактора в i-м опыте, $j = \overline{1,3}$, $i = \overline{1,4}$.

Опыты необходимо проводить по матрице планиро-

вания, отвечающей свойствам, рассмотренным в § 11.2. Из табл. 11.7 (первая строка) видно, что существует матрица эксперимента, отвечающая таким свойствам и имеющая четыре опыта. Эта матрица двухфакторного эксперимента записана в табл. 11.6. Если предположить, что эффекты взаимодействия между факторами отсутствуют, то вектор-столбец X_1X_2 можно использовать для нового фактора X_3. Матрица планирования для этого случая записана в табл. 11.8.

Таблица 11.8

№ опыта	X_0	X_1	X_2	X_3 (X_1X_2)	Y
1	1	+1	+1	+1	y_1
2	1	−1	+1	−1	y_2
3	1	−1	−1	+1	y_3
4	1	+1	−1	−1	y_4

Итак, при исследовании трех факторов по планам полного факторного эксперимента необходимо провести восемь различных опытов. Накладывая ограничение об отсутствии взаимодействия, можно оценить параметры модели (11.4) с помощью четырех опытов (по плану табл. 11.8).

Если анализируется линейная модель с четырьмя факторами $Y=b_0+b_1X_1+b_2X_2+b_3X_3+b_4X_4$, то минимальное число опытов, определяемое количеством оцениваемых параметров (b_0, b_1, b_2, b_3, b_4), равно пяти.

В столбце 3 табл. 11.7 число 5 отсутствует, однако имеется число 8. Матрицу планирования полного факторного эксперимента с восемью опытами можно использовать для расчета модели с четырьмя факторами. В этом случае надо проводить уже не 16 опытов, а восемь. Следовательно, чтобы сократить число опытов, нужно новому фактору присвоить вектор-столбец, принадлежащий взаимодействию, которым можно пренебречь. Тогда значение нового фактора в условиях опытов определяется знаками этого столбца.

Рассмотрим матрицу полного факторного эксперимента для трех факторов, или 2^3 (табл. 11.9).

Проанализируем структуру этой таблицы. Первые три столбца матрицы табл. 11.9 совпадают со столбца-

Таблица 11.9

№ опыта	X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	$X_1X_2X_3$	Y
1	−1	−1	+1	+1	−1	−1	+1	y_1
2	+1	−1	+1	−1	+1	−1	+1	y_2
3	−1	+1	+1	−1	−1	+1	+1	y_3
4	+1	+1	+1	+1	+1	+1	+1	y_4
5	−1	−1	−1	+1	+1	+1	−1	y_5
6	+1	−1	−1	−1	−1	+1	−1	y_6
7	−1	+1	−1	−1	+1	−1	−1	y_7
8	+1	+1	−1	+1	−1	−1	−1	y_8

ми матрицы табл. 11.4 (матрицы планирования для трех факторов). Напомним, что матрица табл. 11.4 была получена (см. §11.2) согласно методу построения матриц планирования. Элементы остальных столбцов найдены соответствующим перемножением первых трех столбцов. Заметим, что при образовании матрицы планирования (табл. 11.4) план 2^2 повторялся дважды, поэтому его называют *полурепликой* полного факторного эксперимента 2^3 и обозначают 2^{3-1} (4 опыта). Полуреплика содержит половину опытов полного факторного эксперимента.

Максимальное количество факторов, которое может быть исследовано с помощью матрицы, записанной в табл. 11.9, равно семи: ($X_1X_2=X_4$; $X_1X_3=X_5$; $X_2X_3=X_6$; $X_1X_2X_3=X_7$). В этом случае четыре фактора (X_4; X_5; X_6; X_7) приравнены к эффектам взаимодействия. Если табл. 11.9 используют для анализа семи факторов, то ее обозначают 2^{7-4} (8 опытов) и называют $^1/_{16}$-репликой полного факторного эксперимента 2^7. План с предельным числом факторов для данной матрицы планирования эксперимента и линейной модели называется *насыщенным*.

На практике редко пользуются даже полурепликой 2^{5-1} (16 опытов), не говоря уже о 2^{6-1} (32 опыта), 2^{7-1} (64 опыта) и т. д. Поэтому с ростом числа факторов возрастает и дробность применяемых реплик. Вопрос о том, какими эффектами взаимодействия можно пренебречь и к какому это приведет риску, должен быть решен до постановки эксперимента по дробным репликам.

§ 11.4. Проведение и обработка результатов эксперимента

Все сведения, необходимые для постановки эксперимента, заносят в специальную таблицу (табл. 11.10).

Таблица 11.10

Наименование	x_1	x_2	...	x_k
Нулевой уровень x_{j0}	x_{10}	x_{20}		x_{k0}
Интервал варьирования h_j	h_1	h_2		h_k
Верхний уровень (+1)	$x_{1в}$	$x_{2в}$		$x_{kв}$
Нижний уровень (−1)	$x_{1н}$	$x_{2н}$		$x_{kн}$

Заметим, что в теории планирования эксперимента, так же как и в дисперсионном анализе, для каждого сочетания факторов проводится не один опыт, а несколько. Такие опыты называют *параллельными*. На практике обычно достаточна постановка двух праллельных опытов. Необходимость в проведении параллельных опытов возникает в том случае, когда исследователь хочет проверить гипотезу об адекватности модели и исследуемого процесса. Эту гипотезу можно проверить, если известна дисперсия воспроизводимости, рассчитываемая по данным параллельных опытов.

После заполнения табл. 11.10 составляют план эксперимента, в который вносят результаты параллельных опытов (табл. 11.11).

Так как невозможно полностью исключить действие внешних факторов, параллельные опыты не дают полно-

Таблица 11.11

№ опыта	X_1	X_2	...	X_k	Параллельные опыты				$\bar{Y}_i$
					Y_1	Y_2	...	Y_m	
1	+1	+1	...	+1	y_{11}	y_{12}	...	y_{1m}	$\bar{y}_1$
2	+1	−1	...	+1	y_{21}	y_{22}	...	y_{2m}	$\bar{y}_2$
...	...	...	...	...	...	...	...	...	...
n	−1	−1	...	+1	y_{n1}	y_{n2}	...	y_{nm}	$\bar{y}_n$

стью совпадающих результатов. Погрешность опыта можно оценить по формуле

$$\hat{s}_i^2 = \frac{\sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2}{m-1}, \quad j = \overline{1, m}, \quad i = \overline{1, n}, \qquad (11.5)$$

где $\hat{s}_i^2$ — *дисперсия воспроизводимости* i-го опыта.

При анализе опытных данных следует использовать критерии математической статистики. Например, резко выделяющиеся значения можно отбрасывать по t-критерию Стьюдента (см. § 6.3), проверять однородность дисперсий $\hat{s}_i^2$ по F-критерию Фишера (см.§ 6.4), производя попарные сравнения. Если дисперсии $\hat{s}_i^2$ однородны, то дисперсия параметра оптимизации

$$\hat{s}^2(y) = \sum_{i=1}^{n} \hat{s}_i^2 / n. \qquad (11.6)$$

Прежде чем приступить к постановке опытов, надо выработать последовательность их проведения. В проведении опытов, запланированных планом эксперимента, рекомендуется случайная последовательность, т. е. необходима *рандомизация* опытов во времени. Поясним сказанное следующим примером.

Пусть требуется поставить опыты по плану, записанному в табл. 11.12.

Таблица 11.12

№ опыта	X_1	X_2	X_3	Y	№ опыта	X_1	X_2	X_3	Y
1	+	+	+	y_1	5	+	+	−	y_5
2	−	−	+	y_2	6	−	−	−	y_6
3	+	−	+	y_3	7	+	−	−	y_7
4	−	+	+	y_4	8	−	+	−	y_8

Допустим также, что экспериментатор может поставить в первый день четыре опыта и во второй — тоже четыре опыта. Ставить опыты в последовательности, записанной в табл. 11.12, нецелесообразно, так как в первых четырех опытах X_3 находится на верхнем уровне, а

в последних — на нижнем, что может вызвать появление систематической ошибки в определении параметра оптимизации (внешние условия совершенно одинаковыми быть не могут). При рандомизации условий эксперимента вероятность такой опасности уменьшается. В рассматриваемом примере необходимо восемь опытов провести в случайной последовательности, для чего используется таблица случайных чисел. В случайном месте таблицы выписывают числа с 1 по 8, отбрасывая числа больше 8 и уже выписанные; например, можно получить такую последовательность: 8, 2, 1, 6, 4, 5, 3, 7. Это значит, что первым следует реализовать опыт № 8, вторым — № 2, третьим — № 1 и т. д.

Если же планируется проведение параллельных опытов, например, по плану табл. 11.15 проводят два параллельных опыта, то необходимо случайно расположить уже 16 чисел. В этом случае также из случайного места таблицы выписывают 16 чисел, например, 2, 15, 9, 5, 12, 14, 8, 13, 16, 1, 3, 7, 4, 6, 11, 10. Вычитая из каждого числа, большего 8, число 8, получаем окончательную последовательность проведения опытов: 2, 7, 1, 5, 4, 6, 8, 5, 8, 1, 3, 7, 4, 6, 3, 2. Следовательно, первым проводят опыт № 2, вторым — № 7 и т. д. Выбранную случайным образом последовательность нарушать не рекомендуется.

После тщательного проведения эксперимента, отбрасывания значений, полученных ошибочно, и проверки однородности дисперсий воспроизводимости переходят к расчету параметров постулируемой модели и ее анализу. В § 11.2 условились в начале эксперимента рассматривать лишь линейные модели на двух уровнях. Как известно из гл. 5, параметры уравнения связи можно отыскать с помощью метода наименьших квадратов. Если функция отклика или уравнение регрессии имеет вид (11.3), то с помощью метода наименьших квадратов необходимо отыскать b_0, b_1, и b_2, для чего можно провести эксперимент по плану табл. 11.5.

Для анализируемой модели с помощью методики, рассмотренной в § 5.6, получаем следующую систему нормальных уравнений:

$$\begin{cases} nb_0 + b_1 \Sigma X_1 + b_2 \Sigma X_2 = \Sigma Y, \\ b_0 \Sigma X_1 + b_1 \Sigma X_1^2 + b_2 \Sigma X_2 X_1 = \Sigma X_1 Y, \\ b_0 \Sigma X_1 + b_1 \Sigma X_1 X_2 + b_2 \Sigma X_2^2 = \Sigma X_2 Y. \end{cases}$$

Подставив в эту систему конкретные значения столбцов X_0, X_1, X_2 из табл. 11.5, имеем ($n=4$):

$$\Sigma X_1 = \Sigma X_2 = 0 \text{ (свойство симметричности)},$$
$$\Sigma X_1^2 = \Sigma X_2^2 = 4 \text{ (свойство нормировки)},$$
$$\Sigma X_1 X_2 = \Sigma X_2 X_1 = 0 \text{ (свойство ортогональности)}.$$

Следовательно, систему нормальных уравнений можно записать в виде

$$\begin{cases} 4b_0 = \Sigma Y, \\ 4b_1 = \Sigma X_1 Y, \\ 4b_2 = \Sigma X_2 Y, \end{cases}$$

откуда $b_0 = \Sigma Y/4$, $b_1 = \Sigma X_1 Y/4$, $b_2 = \Sigma X_2 Y/4$.

Если теперь рассмотреть линейную модель с k факторами, то, рассуждая аналогично, получаем формулу для расчета коэффициентов уравнения регрессии b_j в общем виде:

$$b_j = \sum_{i=1}^{n} \bar{Y}_i X_{ij}/n, \quad j = 0, 1, 2, \ldots, k, \quad i = 1, 2, \ldots, n. \qquad (11.7)$$

Таким образом, вычисления сводятся к умножению столбца Y на столбец соответствующего фактора и алгебраическому сложению полученных значений. Деление результата на число различных опытов в матрице планирования дает искомый коэффициент.

Пример 11.2. Исследуется процесс разделения смеси растворами кислоты. Параметр оптимизации Y — содержание определенного элемента в выходном растворе, %. Факторы: x_1 — концентрация входного раствора, x_2 — концентрация кислоты.

Задача исследования — получение такого сочетания факторов, при котором значение выходного параметра равно 99—100%. Априорные исследования дали возможность построить области определения для каждого из факторов, выбрать нулевой уровень и интервалы варьирования: $0{,}5 < x_1 < 3{,}3 < x_2 < 9$. Исходная информация записана в табл. 11.13.

Таблица 11.13

Наименование	x_1	x_2
Нулевой уровень x_{j0}	1,5	7
Интервал варьирования h_j	0,5	1
Верхний уровень (+1)	2,0	8
Нижний уровень (—1)	1,0	6

Решение. Матрица планирования эксперимента и результаты параллельных опытов (дисперсии воспроизводимости однородны) записаны в табл. 11.14.

Таблица 11.14

№ опыта	X_0	X_1	X_2	$\bar{Y}_i$
1	+1	−1	−1	95
2	+1	+1	−1	90
3	+1	−1	+1	85
4	+1	+1	+1	82

Рассчитаем параметры уравнения связи:

$$b_0 = (95 + 90 + 85 + 82)/4 = 88; \quad b_1 = (-95 + 90 - 85 + 82)/4 = -2{,}0$$

$$b_2 = (-95 - 90 + 85 + 82)/4 = -4{,}5; \quad Y_T = 88 - 2{,}0X_1 - 4{,}5X_2.$$

После того как коэффициенты модели вычислены, решают вопрос о возможности описать полученной моделью изучаемый процесс. Модель, пригодную для описания процесса, называют *адекватной*.

Адекватность модели проверяют по F-критерию Фишера (см. § 6.4), для чего рассчитывают статистику $F = s^2_{ад}/s^2_{(Y)}$, где $s^2_{(Y)}$ — дисперсия параметра оптимизации, или средняя дисперсия воспроизводимости; $s^2_{ад}$ — дисперсия адекватности. Дисперсия адекватности

$$\hat{s}^2_{ад} = \frac{\sum_{i=1}^{n} (\bar{Y}_i - Y_{iT})^2}{f}, \qquad (11.8)$$

причем f равно разности между числом различных опытов и числом параметров уравнения регрессии, а Y_{iT} — значение параметра оптимизации, рассчитанное по уравнению регрессии.

Пример 11.3. Рассчитаем $\hat{s}^2_{ад}$ для примера 11.2 и проверим гипотезу адекватности, если $s^2_{(Y)} = 0{,}625$, $\alpha = 0{,}05$, причем параллельность опытов равна 2.

Решение. Так как $Y_T = 88 - 2{,}0 \cdot X_1 - 4{,}5 X_2$, то

$$Y_{1T} = 88 - 2{,}0(-1) - 4{,}5(-1) = 94{,}5, \quad Y_{2T} = 88 - 2{,}0(+1) - 4{,}5(-1) = 90{,}5;$$

$$Y_{3T} = 88 - 2{,}0(-1) - 4{,}5(+1) = 85{,}5, \quad Y_{4T} = 88 - 2{,}0(+1) - 4{,}5(+1) = 81{,}5;$$

$f = 4 - 3 = 1$, поскольку количество различных опытов равно четырем (параллельность опытов не учитывается), а число оцениваемых параметров (b_0, b_1, b_2) равно трем. Далее имеем:

$$\hat{s}^2_{ад} = (95-94{,}5)^2 + (90-90{,}5)^2 + (85+85{,}5)^2 + (82-81{,}5)^2 =$$

$$= 1; \quad s^2_{(Y)} = 0{,}625; \quad F = 1/0{,}625 \approx 1{,}57.$$

В таблице F-распределения находим $F_{табл}$, соответствующее $\alpha=0{,}05$, $k_1=f=1$, $k_2=n(m-1)=4$. Получаем $F_{табл}=7{,}71$. Так как $1{,}57<7{,}71$, то модель можно считать адекватной.

После установления адекватности модели и исследуемого процесса переходят к проверке значимости отдельных коэффициентов.

При использовании полного факторного эксперимента и дробных реплик погрешности в определении каждого из коэффициентов равны (следствие свойств матрицы планирования § 11.2):

$$\hat{s}^2_{bj} = \hat{s}^2_{(Y)}/n, \tag{11.9}$$

где $\hat{s}^2_{(Y)}$ — средняя дисперсия воспроизводимой: n — число различных опытов.

Значимость коэффициентов обычно проверяют по t-критерию Стьюдента (см. § 6.3), для чего рассчитывают статистику $t=|b_j|/\hat{s}_{b_j}$, которую затем сравнивают с табличным значением $t_{табл}$, взятым из таблицы t-распределения Стьюдента по вероятности проверяемой гипотезы P и количеству степеней свободы, с которым определялась $\hat{s}^2_{(Y)}$.

Пример 11.4. Проверить значимость коэффициентов уравнения $Y_1=88-2{,}0\,X_1-4{,}5\,X_2$, если $p=0{,}95$; $\hat{s}^2_{(Y)}=0{,}625$; количество степеней свободы $s^2_{(Y)}$ равно $k=4$.

Решение. Так как $\hat{s}^2_{(Y)}=0{,}625$, то $\hat{s}^2_{(b_j)}=0{,}625/4$;

$$\hat{s}_{(b_j)} = \sqrt{0{,}625/4} = 0{,}25/2 = 1/8;$$

$$t_1 = 88:1/8 = 88\cdot 8 = 704, \quad t_2 = 2:1/8 = 16,$$

$$t_3 = 9/2:1/8 = (9\cdot 8)/2 = 36; \quad t_{табл} = 2{,}78.$$

Все рассчитанные значения t больше, чем табличное, поэтому все коэффициенты уравнения $Y_1=88-2{,}0\,X_1-4{,}5\,X_2$ значимы.

Доверительные интервалы для каждого из коэффициентов уравнения получают по формуле

$$b_j - t_{табл}\hat{s}_{(Y)}/\sqrt{n} \leqslant \beta_j \leqslant b_j + t_{табл}\hat{s}_{(Y)}/\sqrt{n}.$$

Пример 11.5. Построить доверительный интервал для каждого

коэффициента уравнения $Y_2 = 88 - 2{,}0\,X_1 - 4{,}5\,X_2$, сохранив условия примера 11.4.

Решение. Имеем: $t_{табл} = 2{,}78$, $n = 4$ (число различных опытов), $\hat{s}^2_{(Y)} = 0{,}625$; $t_{табл}\hat{s}_{(Y)}/\sqrt{n} = 2{,}78 \cdot \sqrt{0{,}625}/\sqrt{4} \approx 0{,}35$; $88 - 0{,}35 < \beta_1 < 88 + 0{,}35$; $87{,}65 < \beta_1 < 88{,}35$; $1{,}65 \leqslant \beta_2 \leqslant 2{,}35$; $4{,}15 \leqslant \beta_3 \leqslant 4{,}85$.

Проведенные расчеты дают возможность экспериментатору принять решение о дальнейших исследованиях.

Перевод модели на язык экспериментатора называют *интерпретацией* модели. Задача интерпретации весьма сложна, однако общие рекомендации сводятся к следующему. Сначала устанавливают, в какой мере каждый из факторов влияет на параметр оптимизации. Величина коэффициента регрессии — количественная мера этого влияния. О характере влияния факторов говорят знаки коэффициентов. Знак плюс свидетельствует о том, что с увеличением значения фактора растет величина параметра оптимизации, при знаке минус увеличение значения фактора приводит к уменьшению параметра оптимизации. Если значение параметра оптимизации максимизируется, то увеличение значений всех факторов, коэффициенты которых имеют знак плюс, благоприятно, а знак минус — неблагоприятно. Если же значение параметра оптимизации минимизируется, то следует рассматривать соотношения, противоположные вышеизложенным.

Пример 11.6. Интерпретировать результаты задачи, решаемой в примерах 11.2—11.5.

Решение. Выше установлено, что модель $Y_2 = 88 - 2{,}0\,X_1 - 4{,}5\,X_2$ адекватно описывает исследуемый процесс в выбранных интервалах варьирования факторов. Фактор X_2 (концентрация кислоты) оказывает бóльшее влияние на Y (содержание элемента в выходном растворе, %), чем X_1 (концентрация входного раствора), так как $|4{,}5| > |2{,}0|$. Коэффициенты в уравнении регрессии у обоих факторов имеют знак минус, поэтому уменьшение значений факторов X_1 и X_2 ведет к увеличению параметра оптимизации, а в рассматриваемой задаче параметр Y максимизируется.

Уравнение для натуральных переменных можно получить, используя формулу (11.1). Коэффициенты регрессии изменятся. При этом пропадает возможность интерпретации влияния факторов по величине и знакам коэффициентов регрессии, так как вектор-столбцы натуральных значений переменных в матрице планирования уже не ортогональны, коэффициенты определяют зависимо друг от друга.

Пример 11.7. В задаче, исследуемой в примерах 11.2—11.6, перейти к натуральным переменным.

Решение. Имеем: $Y_2 = 88 - 2{,}0\,X_1 - 4{,}5\,X_2$; $X_1 = (x_1 - 1{,}5)/0{,}5$; $X_2 = (x_2 - 7)/1$; $Y_2 = 88 - 2{,}0(x_1 - 1{,}5)/0{,}5 - 4{,}5(x_2 - 7)/1$; $Y_2 = 88 - 4 \cdot (x_1 - 1{,}5) - 4{,}5(x_2 - 7)$; $Y_2 = 125 - 4x_1 - 4{,}5x_2$.

Глава 12
Основные понятия теории случайных функций

§ 12.1. Понятие о случайной функции

До настоящей главы основным предметом исследования были случайные величины, принимающие те или иные значения в постоянных условиях отдельного опыта. На практике зафиксировать все условия опыта, как правило, невозможно; опыт протекает во времени, в пространстве, при непрерывном действии посторонних причин. Поэтому чаще приходится иметь дело со случайной величиной, непрерывно изменяющейся в процессе опыта. Примерами таких случайных величин могут служить отклонение траектории космического корабля от расчетной в процессе его полета, часовая выработка рабочего-станочника в процессе работы и т. д.

Изменяющаяся в процессе одного опыта случайная величина называется в отличие от обычной случайной величины *случайной функцией*. Приведем другое определение случайной функции. Напомним, что случайной мы называли величину, которая в результате опыта принимала то или иное значение, причем не известно заранее, какое именно. По аналогии, *случайной* назовем функцию, которая в результате опыта может принять тот или иной конкретный вид, причем заранее не известно, какой именно. Конкретный вид, принимаемый случайной функцией в результате опыта, называется *реализацией* случайной функции.

Если над случайной функцией произвести группу повторных опытов, то получим *группу*, или *семейство*, реализаций этой функции.

Пример 12.1. Фактическая траектория космического корабля может отклоняться от расчетной. Это отклонение S колеблется около нулевого уровня и представляет собой случайную функцию времени $S(t)$.

Полет космического корабля можно рассматривать как опыт, в котором случайная функция $S(t)$ принимает определенную реализацию. При каждом следующем полете такого же космического корабля будет получена новая реализация случайной функции. В результате нескольких полетов можно получить семейство реализаций случайной функции $S(t)$ (рис. 12.1).

Случайная функция времени характеризует процесс изменения случайной величины с течением времени. Поэтому случайные функции времени обычно называют *случайными процессами*. Аргументом случайной функции может быть не только время. Например, скорость ветра в атмосфере в данный момент времени — случайная функция трех аргументов — координат точки в пространстве, а при изменении времени скорость — это уже случайная функция четырех аргументов — координат точки пространства и времени.

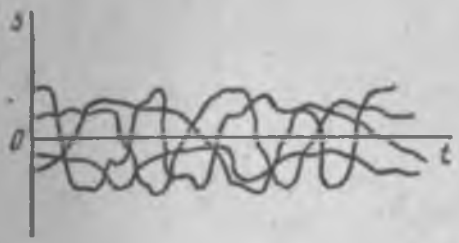

Рис. 12.1

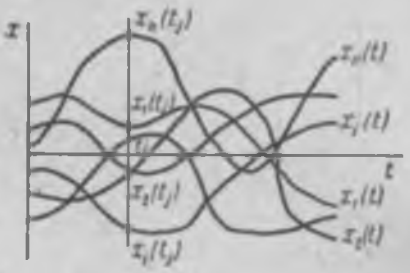

Рис. 12.2

Рассмотрим только случайные функции одного аргумента, обозначив его t. Случайные функции будем обозначать большими буквами латинского алфавита, а их реализации — соответствующими малыми.

Рассмотрим случайную функцию $X(t)$. Пусть над ней произведено n независимых опытов, в результате которых получено n релизаций $x_1(t)$, $x_2(t)$, ..., $x_i(t)$, ..., $x_n(t)$ (рис. 12.2). Каждая реализация есть обычная (неслучайная) функция. Зафиксируем некоторое значение аргумента, например $t=t_j$, и найдем значения n реализаций для данного t_j. Полученные значения $x_1(t_j)$, $x_2(t_j)$, ..., $x_i(t_j)$, ..., $x_n(t_j)$ называют *сечением* n реализацией случайной функции при $t=t_j$. Сечение можно рассматривать как n значений, принятых случайной величиной $X(t_j)$ — ординатой случайной функции при $t=t_j$. Поэтому случайную функцию часто определяют как совокупность случайных величин — ординат функции при различных допустимых значениях ее аргумента.

После интерпретации получившихся результатов переходят к принятию решений о дальнейших исследованиях. Количество возможных ситуаций перечислить невозможно. Остановимся лишь на наиболее часто встречающихся.

Если линейная модель адекватна и коэффициенты регрессии значимы, то можно либо закончить исследования при условии близости оптимума, либо их продолжать. В задаче, рассматриваемой в примерах 11.2—11.6, наибольшее значение параметра оптимизации 95% получено в опыте № 1, в этом случае исследование необходимо продолжить, получив сочетание факторов, при которых содержание элемента в выходном растворе 99—100%. Если линейная модель. адекватна, а часть коэффициентов уравнения регрессии незначима, то можно либо изменить интервалы варьирования факторов, либо отсеить незначимые факторы, произвести параллельные опыты, а если область оптимума близка, закончить исследования.

Заметим, что изменение интервалов варьирования приводит к изменению коэффициентов регрессии. Абсолютные величины коэффициентов регрессии увеличиваются с увеличением интервалов. Инвариантными к изменению интервалов остаются лишь знаки коэффициентов, однако и они могут измениться на противоположные, если при движении «перешагнули» экстремум.

Если линейная модель адекватна, а все коэффициенты уравнения регрессии незначимы (кроме b_0) (чаще всего это происходит вследствие большой ошибки эксперимента или узких интервалов варьирования), необходимо увеличить точность эксперимента, расширить интервалы варьирования. Если область оптимума близка, то можно окончить исследование.

Если линейная модель неадекватна, это значит, что не удается аппроксимировать поверхность отклика плоскостью. В этом случае изменяют интервалы варьирования, выбирают другую точку в качестве нулевого уровня либо используют нелинейную модель.

Если область оптимума близка, то можно окончить исследование.

Особый случай имеет место при использовании насыщенных планов. При значимости всех коэффициентов ничего нельзя сказать об адекватности или неадекватности модели, так как в этом случае невозможно рассчи-

тать $\hat{s}^2_{ад}$ (см. формулу (11.9)) в силу того, что число степеней свободы $f=0$. В этом случае при близости области оптимума можно закончить исследование, в противном случае — продолжить.

Существует множество различных методов продолжения эксперимента до установления оптимальной области. Наиболее старым является метод Гаусса—Зейделя, идея которого сводится к следующему. Все факторы, кроме одного, фиксируют, т. е. продвижение происходит параллельно одной из координатных осей. На этом пути исследователь находит точку наилучшего значения параметра оптимизации, а затем из этой точки двигается параллельно другой оси до тех пор, пока параметр оптимизации не получит запланированного значения. Метод Гаусса — Зейделя требует большого количества опытов. В исследовательской практике широкое распространение получил метод крутого восхождения, предложенный Г. Е. П. Боксом и К. В. Уилсоном.

Метод крутого восхождения дает возможность найти оптимальную область за меньшее число опытов по сравнению с методом Гаусса—Зейделя за счет того, что здесь предусмотрено при переходе от одной точки к другой одновременное изменение значений всех факторов. Выбор последующей точки эксперимента определяется направлением наилучшего изменения параметра оптимизации, т. е. направлением градиента.

Планирование эксперимента достигло в последние годы больших успехов и получило широкое распространение, особенно в лабораторных исследованиях. Однако возможности этого направления в практической деятельности человека используются в настоящее время не полностью. Широкое внедрение методов планирования эксперимента в практику управления производством позволит дать большой экономический эффект. Исследователи, работающие над проблемами теории планирования эксперимента, сосредоточивают свое внимание главным образом на двух направлениях: создание планов с минимальным числом опытов, позволяющих получать уравнения регрессии, достаточно точные для практических целей; создание планов, позволяющих получать высокоточные уравнения регрессии.

Настоящую главу следует рассматривать как введение в планирование эксперимента.

§ 12.2. Законы распределения и основные характеристики случайной функции

Совокупность реализаций случайной функции в какой-то степени характеризует ее свойства. Например, реализации случайных функций $X(t)$ и $Y(t)$, изображенные на рис. 12.3 и 12.4 сплошными тонкими линиями, показывают, что эти функции обладают различными

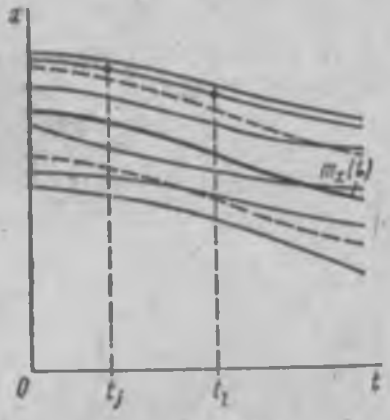

Рис. 12.3

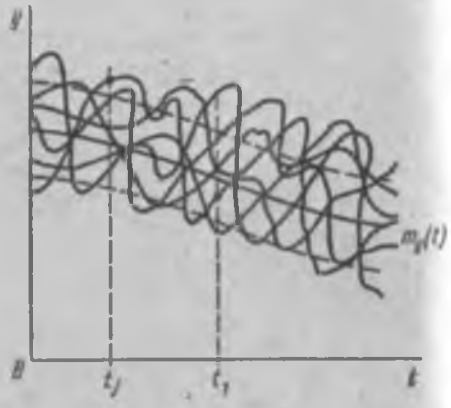

Рис. 12.4

свойствами: для случайной функции $X(t)$ характерно плавное изменение со временем; случайная функция $Y(t)$ имеет резко колебательный характер с беспорядочными колебаниями. Однако подобно тому как вариационный ряд, получаемый в результате серии опытов для определения свойств случайной величины, только приближенно характеризует эту величину, так и совокупность реализаций случайной функции может характеризовать ее свойства только с той или иной степенью приближения.

Будем рассматривать случайную функцию как совокупность случайных величин. В § 3.1 было показано, что любая случайная величина полностью характеризуется законом распределения, при этом наиболее общей формой его задания является функция распределения. Зафиксируем некоторое значение аргумента случайной функции, например $t=t_j$. Обозначим функцию распределения случайной величины $X(t_j)$ через $F[x(t_j)]$. В общем случае при изменении t могут изменяться как параметры функции распределения ординаты, так и ее вид. Придавая t все возможные значения, получаем

множество функций распределения ординат случайной функции.

Одномерной функцией распределения случайной функции $X(t)$ называют такую функцию $F_x(t)$, которая при каждом допустимом значении аргумента случайной функции, например при $t=t_j$, равна функции распределения $F[x(t_j)]$ ординаты $X(t_j)$:

$$F_x(t_j) = F[x(t_j)].$$

Очевидно, что для решения задач, в которых ординаты случайной функции рассматривают изолированно друг от друга, знания одномерной функции распределения вполне достаточно. Однако часто ординаты случайной функции надо рассматривать совместно.

Зафиксируем два значения аргумента t_j и t_l, соответствующие им ординаты обозначим $X(t_j)$ и $X(t_l)$. В § 3.9 было показано, что наиболее общей формой задания двух случайных величин является их совместная функция распределения. Обозначим совместную функцию распределения случайных величин $X(t_j)$ и $X(t_l)$ через $F[x(t_j), x(t_l)]$. Как и в предыдущем случае, при изменении аргументов могут изменяться параметры совместной функции распределения и ее вид.

Двумерной функцией распределения случайной функции $X(t)$ называют такую функцию $F_x(t, t')$, которая при каждой паре допустимых значений аргумента случайной функции, например при $t=t_j$ и $t=t_l$, равна совместной функции распределения $F[x(t_j), x(t_l)]$ ординат $X(t_j)$ и $X(t_l)$:

$$F_x(t_j, t_l) = F[x(t_j), x(t_l)].$$

Напомним, что, зная совместную функцию распределения двух случайных величин, можно определить функцию распределения каждой случайной величины. Поэтому двумерная функция распределения — более полная характеристика случайной функции, чем одномерная. Однако и двумерная функция распределения не является в общем случае полной характеристикой случайной функции, так как в практических задачах приходится рассматривать совместно не две, а гораздо большее число ее ординат.

Продолжая предыдущие рассуждения, можно определить трех-, четырех- и вообще k-мерную функцию распределения случайной функции при любом целом поло-

жительном k. При этом функцию распределения любого низшего порядка можно вычислить по функции распределения высшего порядка. Поэтому каждая следующая функция распределения является более полной характеристикой случайной функции, чем все предыдущие.

Заметим, что при решении практических задач наиболее часто встречаются *нормальные* или *гауссовские* случайные функции. Для таких случайных функций все k-мерные законы распределения (одномерные, двумерные функции распределения и т. д.) являются нормальными. Однако задание случайной функции с помощью k-мерных законов распределения не всегда удобно вследствие их громозкости.

При изучении случайных величин большое значение имеют их основные числовые характеристики: математическое ожидание и дисперсия — для одной случайной величины; математические ожидания, дисперсии и корреляционный момент — для двух случайных величин. Используя эти характеристики, можно решить большое число практических задач. Для случайной функции также вводят простейшие характеристики, аналогичные основным характеристикам случайных величин. Знания их оказывается также достаточно для решения многих задач.

В отличие от характеристик случайных величин, представляющих собой определенные числа, характеристики случайных функций не числа, а функции.

Основными характеристиками случайной функции являются математическое ожидание, дисперсия и корреляционная функция. Определим математическое ожидание случайной функции. Зафиксируем некоторое значение аргумента случайной функции, например $t=t_j$. Математическое ожидание ординаты $X(t_j)$ обозначим $M[X(t_j)]$. При изменении t математическое ожидание ординаты, вообще говоря, изменяется. Придавая t все возможные значения, получаем числовое множество математических ожиданий ординат случайной функции.

Математическим ожиданием случайной функции $X(t)$ называют такую неслучайную функцию $m_x(t)$, которая при каждом допустимом значении аргумента, например при $t=t_j$, равна математическому ожиданию $M[X(t_j)]$ ординаты $X(t_j)$:

$$m_x(t_j) = M[X(t_j)]. \qquad (12.1)$$

По смыслу математическое ожидание случайной функции представляет собой некоторую среднюю функцию, около которой группируются и относительно которой колеблются возможные реализации случайной функции. На рис. 12.3 и 12.4 математические ожидания $m_x(t)$ и $m_y(t)$ случайных функций $X(t)$ и $Y(t)$ изображены жирными линиями.

Аналогично определяется дисперсия случайной функции. *Дисперсией случайной функции* $X(t)$ называют такую неслучайную функцию $D_x(t)$, которая при каждом допустимом значении аргумента, например при $t=t_j$, равна дисперсии $D[X(t_j)]$ ординаты $X(t_j)$:

$$D_x(t_j) = D[X(t_j)]. \qquad (12.2)$$

Очевидно, $D_x(t)$ — неотрицательная функция. Арифметическое значение квадратного корня из дисперсии $D_x(t)$ называют *средним квадратическим отклонением случайной функции:*

$$\sigma_x(t) = \sqrt{D_x(t)}. \qquad (12.3)$$

Среднее квадратическое отклонение случайной функции при каждом t характеризует средний разброс реализаций случайной функции относительно математического ожидания. На рис. 12.3 и 12.4 значения функций $[m(t)-\sigma(t)]$ и $[m(t)+\sigma(t)]$ изображены пунктирными линиями.

Математическое ожидание и дисперсия — важные характеристики случайной функции, но для описания основных особенностей случайных функций этих характеристик недостаточно. Действительно, случайные функции $X(t)$ и $Y(t)$, изображенные на рис. 12.3 и 12.4, имеют примерно одинаковые математические ожидания и дисперсии (средние квадратические отклонения). Однако, как уже отмечалось, характер изменения этих функций различен.

Рассмотрим две ординаты случайной функции $X(t)$, изображенной на рис. 12.3, при значениях аргумента $t=t_j$ и $t=t_l$, т. е. две случайные величины $X(t_j)$ и $X(t_l)$. Очевидно, что при близких значениях t_j и t_l величины $X(t_j)$ и $X(t_l)$ связаны тесной зависимостью: если величина $X(t_j)$ приняла какое-то значение, то и величина $X(t_l)$ с большой вероятностью примет значение, близкое к нему. Очевидно и то, что по мере увеличения промежутка между t_j и t_l зависимость между случайными величинами $X(t_j)$ и $X(t_l)$ затухает очень медленно.

Напротив, для случайной функции $Y(t)$, изображенной на рис. 12.4, ординаты $Y(t_j)$ и $Y(t_l)$ даже при близких значениях t_j и t_l практически не зависят друг от друга. Для такой случайной функции характерно быстрое затухание зависимости между ее ординатами при увеличении расстояния между ними.

Известно, что степень зависимости двух случайных величин может быть охарактеризована корреляционным моментом, или ковариацией. Для характеристики степени зависимости между двумя ординатами случайной функции, относящимися к различным значениям аргумента t, используют корреляционную функцию, которую определяют следующим образом.

Пусть $X(t_j)$ и $X(t_l)$ — две ординаты случайной функции $X(t)$, а $K[X(t_j), X(t_l)]$ — их корреляционный момент. Напомним, что он равен математическому ожиданию произведения этих случайных величин, предварительно центрированных относительно их математических ожиданий, т. е. $K[X(t_j), X(t_l)] = M[\mathring{X}(t_j), \mathring{X}(t_l)]$, где $\mathring{X}(t_j) = X(t_j) - M[X(t_j)]$, а $\mathring{X}(t_l) = X(t_l) - M[X(t_l)]$.

Корреляционной функцией случайной функции $X(t)$ называют такую неслучайную функцию $k_x(t, t')$, которая при каждой паре допустимых значений аргумента t, например при $t = t_j$ и $t = t_l$, равна корреляционному моменту $K[X(t_j), X(t_l)]$ ординат $X(t_j)$ и $X(t_l)$:

$$k_x(t_j, t_l) = K[X(t_j), X(t_l)] = M[\mathring{X}(t_j), \mathring{X}(t_l)]. \qquad (12.4)$$

Так как корреляционный момент случайных величин $X(t_j)$ и $X(t_l)$ не зависит от последовательности, в которой эти величины рассматриваются, то и корреляционная функция не меняется при перемене аргументов местами: $k_x(t, t') = k_x(t', t)$.

Вернемся к случайным функциям $X(t)$ и $Y(t)$, изображенным на рис. 12.3 и 12.4. Очевидно, что их корреляционные функции совершенно различны. Так как по мере увеличения промежутка между t_j и t_l зависимость между ординатами случайной функции $X(t)$ затухает очень медленно, то и ее корреляционная функция $k_x(t, t')$ с ростом промежутка (t, t') убывает медленно; напротив, корреляционная функция $k_y(t, t')$ случайной функции $Y(t)$ с увеличением промежутка (t, t') убывает быстро.

Выясним, что происходит с корреляционной функцией $k_x(t, t')$, когда ее аргументы совпадают. Пусть $t=t'=t_j$. Учитывая формулы (12.4) и (12.2), можно записать следующую цепочку равенств:

$$k_x(t_j, t_j) = K[X(t_j), X(t_j)] = M[\mathring{X}(t_j), \mathring{X}(t_j)] = M[\mathring{X}(t_j)]^2 = \\ = M\{X(t_j) - M[X(t_j)]\}^2 = D[X(t_j)] = D_x(t_j),$$

т. е. корреляционная функция при $t=t'=t_j$ равна дисперсии ординаты $X(t_j)$. Придавая аргументу t все возможные значения и полагая при этом, что $t=t'$, получаем $k_x(t, t)=D_x(t)$, т. е. при $t=t'$ корреляционная функция является дисперсией случайной функции.

Таким образом, необходимость в дисперсии как отдельной характеристики случайной функции отпадает. В качестве основных характеристик случайной функции достаточно рассмотреть ее математическое ожидание и корреляционную функцию.

Вместо корреляционной функции $k_x(t, t')$ можно использовать нормированную корреляционную функцию

$$\rho_x(t, t') = \frac{k_x(t, t')}{\sqrt{D_x(t)}\sqrt{D_x(t')}}, \qquad (12.5)$$

которая при каждой паре допустимых значений аргумента случайной функции, например при $t=t_j$ и $t=t_l$, равна нормированному корреляционному моменту или коэффициенту корреляции ординат $X(t_j)$ и $X(t_l)$:

$$\rho_x(t_j, t_l) = \frac{k_x(t_j, t_l)}{\sqrt{D_x(t_j)}\sqrt{D_x(t_l)}} = \frac{K[X(t_j), X(t_l)]}{\sqrt{D[X(t_j)]}\sqrt{D[X(t_l)]}} = \\ = \rho[X(t_j), X(t_l)]. \qquad (12.6)$$

Заметим, что математическое ожидание и корреляционная функция являются исчерпывающими характеристиками только для нормальной случайной функции. И вот почему. В соответствии с определением нормальной случайной функции все ее k-мерные функции распределения, т. е. совместные законы распределения любых ее k ординат (k случайных величин), являются нормальными. В § 3.4 было показано, что нормальный закон распределения одной случайной величины полностью определяется ее математическим ожиданием и дисперсией. В § 3.10 показано, что нормальный закон распределения двух случайных величин полностью опреде-

ляется математическими ожиданиями, дисперсиями и коэффициентом корреляции. Можно доказать, что и нормальный закон распределения большего числа случайных величин определяется их математическими ожиданиями, дисперсиями и парными коэффициентами корреляции. Поэтому для определения нормальной случайной функции достаточно знать ее математическое ожидание, дисперсию и нормированную корреляционную функцию или математическое ожидание и корреляционную функцию (так как, зная дисперсию и нормированную корреляционную функцию, можно восстановить корреляционную функцию и, наоборот, зная корреляционную функцию, можно определить дисперсию и нормированную корреляционную функцию).

Изучая случайные функции, кроме математического ожидания и корреляционной функции можно рассматривать, например, начальные и центральные моменты случайной функции и другие характеристики. Способы их задания аналогичны способам задания основных характеристик.

§ 12.3. Определение основных характеристик случайной функции из опыта

Пусть над случайной функцией $X(t)$ проведено n независимых опытов и в результате получено n реализаций случайной функции (рис. 12.5). Найдем оценки ха-

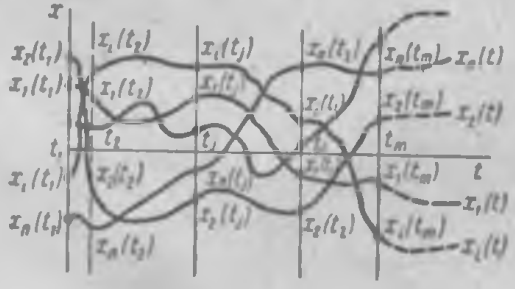

Рис. 12.5

рактеристик случайной функции: математического ожидания $m_x(t)$, дисперсии $D_x(t)$ и корреляционной функции $k_x(t, t')$. Для наглядности последующих рассуждений предположим, что аргумент случайной функции — время.

Промежуток времени $[0, T]$, в течение которого проводят наблюдения, разобьем на интервалы точками $0=$

$=t_1, t_2, \ldots, t_m=T$. Проводя каждый опыт над случайной функцией, будем регистрировать значения, принятые ею в моменты времени $t_1, t_2, \ldots, t_m$. Зарегистрированные значения поместим в табл. 12.1, число строк которой равной числу реализации n, а число столбцов равно m. В таблице в i-й строке и j-м столбце помещено $x_i(t_j)$ —

Таблица 12.1

Номер опыта	Аргумент / Ордината / Реализация	t_1	t_2	...	t_j	...	t_l	...	t_m
		$X(t_1)$	$X(t_2)$	...	$X(t_j)$	...	$X(t_l)$	...	$X(t_m)$
1	$x_1(t)$	$x_1(t_1)$	$x_1(t_2)$	...	$x_1(t_j)$	...	$x_1(t_l)$	...	$x_1(t_m)$
2	$x_2(t)$	$x_2(t_1)$	$x_2(t_2)$	...	$x_2(t_j)$	...	$x_2(t_l)$	...	$x_2(t_m)$
...	...	...	...	...	...	...	...	...	...
i	$x_i(t)$	$x_i(t_1)$	$x_i(t_2)$	...	$x_i(t_j)$	...	$x_i(t_l)$	...	$x_i(t_m)$
...	...	...	...	...	...	...	...	...	...
n	$x_n(t)$	$x_n(t_1)$	$x_n(t_2)$	...	$x_n(t_j)$	...	$x_n(t_l)$	...	$x_n(t_m)$

значение случайной функции, которое наблюдалось в i-м опыте в момент времени t_j. Полученный материал не что иное, как результаты n опытов над каждой из случайных величин $X(t_1), X(t_2), \ldots, X(t_m)$.

Рассмотрим случайную величину $X(t_j)$. Из § 5.3 известно, что оценкой ее математического ожидания $M[X(t_j)]$ является средняя арифметическая $\bar{x}(t_j)$ результатов наблюдений:

$$M[X(t_j)] \approx \bar{x}(t_j) = \left(\sum_{i=1}^{n} x_i(t_j)\right)/n, \qquad (12.7)$$

а оценкой ее дисперсии $D[X(t_j)]$ является исправленная выборочная дисперсия $\hat{s}_x^2(t_j)$:

$$D[X(t_j)] \approx \hat{s}_x^2(t_j) = \left(\sum_{i=1}^{n} [x_i(t_j) - \bar{x}(t_j)]^2\right)/(n-1). \qquad (12.8)$$

Аналогично найдем оценки математических ожиданий и дисперсий всех случайных величин $X(t_1)$, $X(t_2)$, ..., $X(t_m)$. Получаем два ряда: ряд значений средних арифметических $X(t_1), X(t_2), ..., X(t_m)$, используя который можно установить зависимость $\bar{X}(t)$ между временем и средними арифметическими ординат; ряд значений исправленных выборочных дисперсий $\hat{s}_x^2(t_1)$, $\hat{s}_x^2(t_2)$, ..., $\hat{s}_x^2(t_m)$, с помощью которого можно установить зависимость $s_x^2(t)$ между временем и выборочными дисперсиями ординат. Зависимости $\bar{x}(t)$ и $\hat{s}_x^2(t)$ называют соответственно *оценками математического ожидания* $m_x(t)$ *и дисперсии* $D_x(t)$ *случайной функции* $X(t)$. Эти зависимости можно аппроксимировать каким-либо аналитическим выражением, используя, например, метод наименьших квадратов (см. § 5.6).

Рассмотрим две случайные величины $X(t_j)$ и $X(t_l)$. Оценкой их корреляционного момента $K[X(t_j), X(t_l)]$ является исправленный выборочный корреляционный момент, который обозначим $\hat{k}_x(t_j, t_l)$:

$$K[X(t_j), X(t_l)] \approx \hat{k}_x(t_j, t_l) = \frac{\sum_{i=1}^{n} [x_i(t_j) - \bar{x}(t_j)][x_i(t_l) - \bar{x}(t_l)]}{n-1}. \quad (12.9)$$

Вычислим выборочные корреляционные моменты для всех возможных пар из m случайных величин $X(t_1)$, $X(t_2)$, ..., $X(t_m)$ и запишем их в табл. 12.2, число строк и столбцов которой равно m.

Для любого момента времени t_j, где $j=1, 2, ..., m$, справедлива следующая цепочка равенств:

$$\hat{k}_x(t_j, t_j) = \frac{\sum_{i=1}^{n} [x_i(t_j) - \bar{x}(t_j)][x_i(t_j) - \bar{x}(t_j)]}{n-1} = \hat{s}^2(t_j),$$

поэтому по главной диагонали таблицы расположены исправленные выборочные оценки дисперсий рассматриваемых ординат. Клетки под главной диагональю не заполнены, так как из формулы (12.9) следует, что $\hat{k}_x(t_j, t_l) = \hat{k}_x(t_l, t_j)$.

Таблица 12.2

t'	t	t_1	t_2	...	t_j	...	t_l	...	t_m
	ордината	$X(t_1)$	$X(t_2)$	...	$X(t_j)$	...	$X(t_l)$	...	$X(t_m)$
t_1	$X(t_1)$	$\hat{s}_x^2(t_1)$	$\hat{k}_x(t_1, t_2)$	...	$\hat{k}_x(t_1, t_j)$	...	$\hat{k}_x(t_1, t_l)$	...	$\hat{k}_x(t_1, t_m)$
t_2	$X(t_2)$		$\hat{s}_x^2(t_2)$	...	$\hat{k}_x(t_2, t_j)$	...	$\hat{k}_x(t_2, t_l)$	...	$\hat{k}_x(t_2, t_m)$
...	...			...	...	...	...	...	...
t_j	$X(t_j)$				$\hat{s}_x^2(t_j)$	...	$\hat{k}_x(t_j, t_l)$	...	$\hat{k}_x(t_j, t_m)$
...	...					...	...	...	...
t_l	$X(t_l)$						$\hat{s}_x^2(t_l)$	...	$\hat{k}_x(t_l, t_m)$
...								...	...
t_m	$X(t_m)$								$\hat{s}_x^2(t_m)$

Используя эту таблицу, можно установить зависимость $\hat{k}_x(t, t')$ между возможными парами моментов времени t и t' и выборочными корреляционными моментами соответствующих ординат, которая и называется *оценкой корреляционной функции* $k_x(t, t')$ *случайной функции* $X(t)$.

За оценку нормированной корреляционной функции $\rho_x(t, t')$ принимают зависимость $r_x(t, t')$ между возможными парами моментов времени t и t' и выборочными нормированными корреляционными моментами, или, иначе говоря, выборочными коэффициентами корреляции соответствующих ординат. Напомним, что выборочный коэффициент корреляции случайных величин $X(t_j)$ и $X(t_l)$ находят по формуле

$$r_x(t_j, t_l) = \frac{\sum_{i=1}^{n} [x_i(t_j) - \bar{x}(t_j)][x_i(t_l) - \bar{x}(t_l)]}{\sqrt{\sum_{i=1}^{n} [x_i(t_j) - \bar{x}(t_j)]^2} \sqrt{\sum_{i=1}^{n} [x_i(t_l) - \bar{x}(t_l)]^2}} =$$

$$= \frac{\hat{k}_x(t_j, t_l)}{\sqrt{\hat{s}^2(t_j)}\sqrt{\hat{s}^2(t_l)}}. \qquad (12.10)$$

При нахождении оценок для характеристик случайной функции регистрировались ее значения в моменты времени $t_1, t_2, \ldots, t_m$. Обычно эти моменты времени задают равноотстоящими; при этом величину интервала между соседними моментами выбирают в зависимости от вида полученных в результате опытов реализаций так, чтобы по значениям ординат в эти моменты можно было восстановить в основном вид реализаций. Точность оценок характеристик случайной функции тем выше, чем больше проведено опытов над ней.

Пример 12.2. В результате обработки детали на ее поверхности остаются неровности, являющиеся результатом неравномерности процесса резания, пластических деформаций детали и других причин, повторяющихся нерегулярно. Если сделать сечения детали вдоль обрабатываемой поверхности, то кривые профиля отличаются друг от друга. Ординату профильной кривой можно рассматривать как случайную функцию $X(t)$ протяженности трассы обработки поверхности t, мм. На рис. 12.6 сплошными тонкими линиями изображены пять профильных кривых или, иначе говоря, пять реализаций случайной функции $X(t)$. Требуется оценить $m_x(t)$, $D_x(t)$, $k_x(t, t')$ и $\rho_x(t, t')$. Заметим, что мы ограничились пятью реализа-

циями только потому, что пример носит иллюстративный характер.

Р е ш е н и е. Из рис. 12.6 следует, что случайная функция $X(t)$ меняется сравнительно плавно. Поэтому сечения можно брать не очень часто, например через 50 мм. Тогда случайная функция сведется к системе случайных величин, взятых при $t=0, 50, ..., 350$. Будем измерять значения этих случайных величин, или, иначе говоря, ординату профиля, передвигаясь вдоль каждой профильной кривой. Результаты замеров приведены в табл. 12.3.

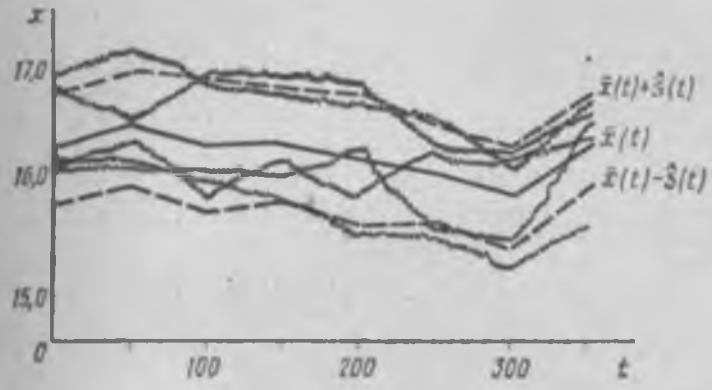

Рис. 12.6

В таблице, например, число 15,82 мм, стоящее на пересечении 2-й строки и 3-го столбца, есть значение ординаты 2-й профильной кривой при протяженности трассы обработки (абсциссе), равной 100 мм.

Найдем оценки для характеристик случайных величин $X(0)$, $X(50), ..., X(350)$. Суммируя значения по столбцам и деля сумму

Таблица 12.3

Номер профильной кривой	Номер сечения	1	2	3	4
	Протяженность трассы, мм	0	50	100	150
	Ордината / Реализация	$X(0)$	$X(50)$	$X(100)$	$X(150)$
1	$x_1(t)$	16,18	16,16	15,97	15,77
2	$x_2(t)$	16,15	16,46	15,82	16,13
3	$x_3(t)$	16,70	16,50	16,90	16,89
4	$x_4(t)$	16,11	16,10	16,09	16,03
5	$x_5(t)$	16,93	17,18	16,73	16,68

Продолжение

Номер профильной кривой	Номер сечения	5	6	7	8
	Протяженность трассы, мм	200	250	300	350
	Ордината / Реализация	X (200)	X (250)	X (300)	X (360)
1	$x_1(t)$	15,56	15,52	15,29	15,61
2	$x_2(t)$	15,80	16,12	16,15	16,33
3	$x_3(t)$	16,70	16,30	16,20	16,50
4	$x_4(t)$	16,22	15,58	15,44	16,42
5	$x_5(t)$	16,63	16,50	16,08	16,58

на число реализаций $n=5$, найдем приближенно зависимость средних арифметических от протяженности трассы t (табл. 12.4).

Таблица 12.4

t	0	50	100	150	200	250	300	350
$\bar{x}(t)$	16,30	16,48	16,30	16,30	16,18	16,03	15,84	16,29

На рис. 12.6 оценка $\bar{x}(t)$ математического ожидания профиля поверхности показана жирной линией.

Далее находим оценки для дисперсий ординат $X(0)$, $X(50)$, ..., $X(350)$ и их корреляционных моментов. Вычисления проведем в следующем порядке. Сначала вычислим исправленные выборочные дисперсии ординат по формуле (12.8), или, иначе говоря, выборочные корреляционные моменты таких ординат $X(t_j)$ и $X(t_i)$, для которых $t_j=t_i$. Найденные дисперсии расположены на главной диагонали табл. 12.5. Затем, используя формулу (12.9), вычислим исправленные выборочные корреляционные моменты ординат, для которых $t_i-t_j=50$ мм, расположим их на первой параллели главной диагонали табл. 12.5. На второй параллели расположим оценки корреляционных моментов ординат, для которых $t_i-t_j=100$ мм и т. д.

Зависимость выборочных оценок корреляционных моментов $k_x(t, t')$ ординат $X(t)$ и $X(t')$ от аргументов t и t', представленная в табл. 12.5, дает оценку корреляционной функции $k_x(t, t')$ профиля поверхности.

Таблица 12.5

t' \ t	0	50	100	150	200	250	300	350
0	0,1575	0,1406	0,1656	0,1539	0,1538	0,1304	0,0976	0,0794
50		0,1844	0,1260	0,1383	0,1290	1,1662	0,1242	0,0936
100			0,2321	0,2076	0,2210	0,1339	0,1089	0,1070
150				0,2188	0,2150	0,1743	0,1545	0,1381
200					0,2508	0,1444	0,1258	0,1616
250						0,1981	0,1823	0,1222
300							0,1864	0,1199
350								0,1522

Зависимость исправленных выборочных оценок дисперсий ординат и оценок их средних квадратических отклонений от протяженности трассы обработки t, представленная в табл. 12.6, дает оценку дисперсии $D_x(t)$ и среднего квадратического отклонения $\sigma_x(t)$ профиля поверхности.

Таблица 12.6

t	0	50	100	150
$\hat{s}_x^2(t)$	0,1575	0,1844	0,2321	0,2188
$\hat{s}_x(t)$	0,396	0,429	0,482	0,467
$\bar{x}(t)+\hat{s}_x(t)$	16,696	16,909	16,782	16,767
$\bar{x}(t)-\hat{s}_x(t)$	15,904	16,051	15,818	15,833

t	200	250	300	350
$\hat{s}_x^2(t)$	0,2508	0,1981	0,1864	0,1522
$\hat{s}_x(t)$	0,501	0,445	0,432	0,389
$\bar{x}(t)+\hat{s}_x(t)$	16,681	16,475	16,272	16,679
$\bar{x}(t)-\hat{s}_x(t)$	15,679	15,585	15,408	15,901

Для характеристики степени разбросанности реализаций случайной функции вокруг оценки ее математического ожидания на рис. 12.6 пунктирными линиями изображены приведенные в табл. 12.6 значения функций $\overline{x}(t)+\hat{s}_x(t)$ и $\overline{x}(t)-\hat{s}_x(t)$.

Разделив значения корреляционной функции $\hat{k}_x(t, t')$, помещенные в табл. 12.5, на произведения соответствующих значений среднего квадратического отклонения $\hat{s}_x(t)$, взятых из табл. 12.6, получим таблицу значений нормированной корреляционной функции $r_x(t, t')$ (табл. 12.7). Например,

$$r_x(200,50)=\hat{k}_x(200,50)/(\hat{s}_x(200)\cdot\hat{s}_x(50)=0,1290/(0,501\cdot0,429)\approx$$
$$\approx 0,60.$$

Таблица 12.7

t' \ t	0	50	100	150	200	250	300	350
0	1,00	0,82	0,87	0,83	0,77	0,73	0,57	0,51
50		1,00	0,60	0,68	0,60	0,80	0,67	0,56
100			1,00	0,92	0,91	0,62	0,52	0,57
150				1,00	0,91	0,83	0,76	0,76
200					1,00	0,65	0,58	0,82
250						1,00	0,94	0,70
300							1,00	0,71
350								1,00

§ 12.4. Понятие о стационарной случайной функции

На практике очень часто встречаются случайные функции, изменяющиеся во времени примерно однородно, имеющие вид непрерывных случайных колебаний около некоторого постоянного среднего значения, причем ни средняя амплитуда, ни характер этих колебаний с течением времени не обнаруживает существенных изменений. К таким функциям можно отнести, например, колебания напряжений в электрической осветительной сети, случайные шумы в радиоприемнике и т. д. Как правило, математическое ожидание такой случайной функции постоянно и не зависит от значения аргумента, а корреляционная функция зависит только от разности

аргументов $\Delta t = t' - t$ и не зависит от самих значений аргументов, т. е. выполняются следующие соотношения:

$$m_x(t) = m_x = \text{const}, \tag{12.11}$$

$$k_x(t, t') = k_x(t' - t) = k_x(\Delta t). \tag{12.12}$$

Случайную функцию, математическое ожидание которой постоянно, а корреляционная функция зависит только от разности аргументов, называют *стационарной в широком смысле.*

Так как дисперсия случайной функции равна ее корреляционной функции при $t' = t$, то, учитывая соотношение (12.12), запишем следующую цепочку равенств:

$$D_x(t) = k_x(t, t) = k_x(t - t) = k_x(0) = D_x = \text{const}, \tag{12.13}$$

т. е. дисперсия функции, стационарной в широком смысле, как и ее математическое ожидание, постоянна и не зависит от значения аргумента.

В приведенное выше определение стационарности функции включены слова «в широком смысле», поскольку, вообще говоря, случайная функция $X(t)$ называется стационарной относительно каких-либо характеристик, если эти характеристики не изменяются при любом сдвиге аргументов, от которых они зависят по оси t, иначе говоря, определенные характеристики инвариантны относительно любых сдвигов по t. В практических задачах случайные функции могут иметь различное количество характеристик, инвариантных относительно сдвигов вдоль оси времени, поэтому можно говорить об их стационарности «в большей или меньшей степени». Еще раз напомним, что рассмотрение t в качестве времени носит условный характер, аргумент случайной функции может иметь и другую природу.

Стационарность в широком смысле представляет простейший вид стационарности. Для случайной функции, стационарной в широком смысле, только основные ее характеристики (математическое ожидание и корреляционная функция) не изменяются при любом сдвиге аргументов, от которых они зависят. Действительно, равенство (12.11) справедливо при всех допустимых значениях аргумента t, поэтому при сдвиге аргумента по оси t на промежуток Δt $m_x(t + \Delta t) = m_x = m_x(t)$, т. е. математическое ожидание инвариантно относительно любо-

го сдвига аргумента по оси t. Равенство (12.12) справедливо при всех допустимых значениях аргументов t и t'. Дадим обоим аргументам приращение Δt. Тогда $k_x(t+\Delta t,\ t'+\Delta t)=k_x(t'+\Delta t-t-\Delta t)=k_x(t'-t)=k_x(t,\ t')$, т. е. и корреляционная функция инвариантна относительно любого сдвига аргументов по оси t. Требование стационарности в широком смысле накладывает на случайную функцию наименьшие ограничения.

Другим крайним случаем является полная стационарность или стационарность в узком смысле, когда все

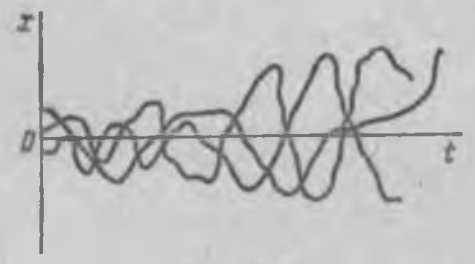

Рис. 12.7

без исключения вероятностные характеристики случайной функции инвариантны относительно произвольных сдвигов по оси независимой переменной.

Заметим, что если функция стационарна в узком смысле, то она стационарна и в широком смысле. Обратное утверждение верно только для нормальной случайной функции, так как она полностью определяется математическим ожиданием и корреляционной функцией. В дальнейшем, говоря о стационарных случайных функциях, будем иметь в виду только стационарные в широком смысле случайные функции. Для краткости такие случайные функции условимся называть просто стационарными.

Нестационарная функция имеет либо изменяющееся с течением времени математическое ожидание, т. е. нарушается равенство (12.11), либо ее корреляционная функция изменяется при сдвиге аргументов по оси t, т. е. нарушается равенство (12.12), либо нарушаются оба условия одновременно. Например, на рис. 12.3 изображена функция, математическое ожидание которой изменяется с течением времени (про такую функцию говорят, что она имеет тенденцию развития во времени). На

рис. 12.7 изображена функция с постоянным математическим ожиданием, но с изменяющейся с течением времени дисперсией, о чем свидетельствует увеличение с ростом аргумента отклонения реализаций случайной функции от математического ожидания.

Заметим, что если случайная функция $X(t)$ нестационарна только за счет переменного математического ожидания $m_x(t)$, то вместо $X(t)$ следует рассмотреть случайную функцию $\overset{\circ}{X}(t)=X(t)-m_x(t)$, которая удовлетворяет условиям стационарности. Действительно, ее математическое ожидание постоянно и равно нулю: $m_{\overset{\circ}{x}}(t) = M[X(t)-m_x(t)] = M[X(t)]-M[m_x(t)]= m_x(t)-m_x(t)=0$; далее, в силу того, что корреляционный момент двух случайных величин совпадает с корреляционным моментом этих случайных величин, предварительно центрированных относительно их математических ожиданий $(K[X(t),X(t')]=M\{[X(t)-M(X(t))]\times \times[X(t')-M(X(t'))]\}= M\{\overset{\circ}{X}(t)\overset{\circ}{X}(t')\} =K[\overset{\circ}{X}(t),\overset{\circ}{X}(t')])$, корреляционная функция $k_{\overset{\circ}{x}}(t, t')$ совпадает с корреляционной функцией $k_x(t, t')$ и, следовательно, как и $k_x(t, t')$, зависит только от разности аргументов: $k_{\overset{\circ}{x}}(t, t')= =k_x(t, t')=k_x(t'-t)=k_x(\Delta t)$.

Корреляционная функция любой случайной функции не меняет значения при перестановке значений аргументов, поэтому для стационарной случайной функции, учитывая соотношение (12.12), можно записать следующую цепочку равенств: $k_x(\Delta t)=k_x(t'-t)=k_x(t, t')=k_x(t', t)=k_x(t-t')=k_x(-\Delta t)$. Таким образом, $k_x(\Delta t)= =k_x(-\Delta t)$, т. е. корреляционная функция стационарной случайной функции $X(t)$ — четная функция разности аргументов.

На практике вместо корреляционной функции $k_x(\Delta t)$ часто используют нормированную корреляционную функцию

$$\rho_x(\Delta t)=k_x(\Delta t)/D_x. \qquad (12.14)$$

Так как для стационарной случайной функции $D_x= =\text{const}$, то $\rho_x(\Delta t)$ также четная функция, т. е. $\rho_x(\Delta t)= =\rho_x(-\Delta t)$. По смыслу $\rho_x(\Delta t)$ при заданном Δt есть коэффициент корреляции между ординатами случайной функции, разделенными по оси независимой переменной интервалом Δt.

§ 12.5. Определение характеристик стационарной случайной функции из опыта

Прежде чем находить оценки характеристик стационарной случайной функции из опыта, выясним, можно ли, имея результаты наблюдений над случайной функцией, ответить на вопрос: стационарна она или нет? В § 12.3 показано, как найти оценки характеристик любой случайной функции. В примере 12.2 по результатам обработки наблюдений были получены в табличной форме зависимости $\bar{x}(t)$: $\hat{s}_x^2(t)$, $\hat{k}_x(t, t')$ и $r_x(t, t')$ (см. табл. 12.4—12.7), которые и являются оценками основных характеристик $m_x(t)$, $D_x(t)$, $k_x(t, t')$, $\rho_x(t, t')$ случайной функции $X(t)$.

В табл. 12.4 и 12.6, по существу, приведены значения средних арифметических и исправленных выборочных дисперсий для набора случайных величин — ординат случайной функции. Используя теорию проверки статистических гипотез, можно выяснить, случайны или нет различия средних арифметических ординат и их исправленных выборочных дисперсий. Если эти различия случайны, то средние арифметические ординат являются оценками одного и того же математического ожидания m_x, а выборочные дисперсии — оценками одной и той же дисперсии D_x, т. е. случайная функция $X(t)$ имеет постоянные математическое ожидание и дисперсию.

В табл. 12.7 на первой параллели главной диагонали (коднагонали) расположены выборочные коэффициенты корреляции между ординатами, разделенными интервалом 50 мм. Используя теорию проверки статистических гипотез, можно выяснить, случайны или нет различия этих коэффициентов корреляции. Если различия случайны, то эти коэффициенты являются оценками одного и того же коэффициента корреляции ρ_x (50). Аналогично можно проверить, являются ли выборочные коэффициенты корреляции, расположенные на второй кодиагонали табл. 12.7, оценками одного и того же коэффициента корреляции ρ_x (150), и т. д. Если окажется, что различия выборочных коэффициентов корреляции по каждой диагонали вызваны случайными причинами, то можно сделать вывод, что коэффициент корреляции любых двух ординат случайной функции не зависит от значений аргументов, при которых были взя-

ты ординаты, а зависит только от разности аргументов. Такой способ проверки стационарности функции довольно трудоемок и не всегда правомерен. Дело в том, что при решении практических задач, как правило, исследователь ограничен небольшим числом реализаций случайной функции. Это может привести к тому, что теория статистической проверки гипотез, предполагающая достаточно большое число наблюдений, не подтвердит гипотезы о стационарности функции, хотя природа явления позволяет судить о том, что оно должно описываться стационарной случайной функцией. Поэтому решить, стационарна или нет исследуемая функция, можно только изучив внутреннюю структуру явления.

Со стационарными случайными функциями очень часто приходится сталкиваться на практике, особенно при изучении физических и технических явлений. Как отмечалось выше, стационарные случайные функции характеризуются прежде всего однородностью протекания во времени. Этот признак и является основным при определении, стационарна исследуемая функция или нет. Как правило, случайный процесс в любой динамической системе начинается с нестационарной стадии — с так называемого «переходного процесса». После затухания «переходного процесса» система обычно переходит на установившийся режим и тогда случайные процессы, протекающие в ней, могут быть описаны стационарными функциями.

Покажем, как по результатам наблюдений над случайной функцией оценить ее характеристики, предполагая, что она стационарная.

Пример 12.3. Рассматривая ординату профильной кривой $X(l)$ (см. пример 12.2) как стационарную случайную функцию, найти оценки ее математического ожидания, дисперсии и нормированной корреляционной функции.

Решение. При вычислении характеристик $X(l)$ в примере 12.2 стационарность функции не предполагалась. Если судить по полученным оценкам, то можно прийти к выводу, что случайная функция стационарной не является: оценки математических ожиданий (табл. 12.4) и дисперсий ординат (табл. 12.6) меняются со временем. Оценки корреляционных моментов, как и оценки коэффициентов корреляции, расположенные вдоль параллелей главной диагонали (см. табл. 12.5 и 12.7), не постоянны. Однако, принимая во внимание весьма ограниченное число реализаций и в связи с этим наличие большого элемента случайности в найденных оценках, эти отступления от стационарности вряд ли можно считать существенными, тем более что они не носят сколько-нибудь закономерного

характера. Поэтому целесообразна замена $X(t)$ стационарной функцией.

Для получения оценки ее математического ожидания вычислим среднюю из средних арифметических ординат случайной функции, приведенных в табл. 12.4:

$$m_x \approx \bar{\bar{x}} = (\bar{x}(0) + \bar{x}(50) + \ldots + \bar{x}(350))/8 = 15{,}84.$$

Осреднив оценки дисперсий, приведенные в табл. 12.6, получим оценку для дисперсии стационарной случайной функции:

$$D_x \approx \hat{s}_x^2 = \left(\hat{s}^2(0) + \hat{s}^2(50) + \ldots + \hat{s}^2(350)\right)/8 = 0{,}1975.$$

Оценка среднего квадратического отклонения профиля поверхности

$$\hat{s}_x = \sqrt{\hat{s}_x^2} = \sqrt{0{,}1975} \approx 0{,}44.$$

Вычислим нормированную корреляционную функцию, по-прежнему предполагая, что $X(t)$ — стационарная функция. Для стационарной случайной функции корреляционная функция зависит только от разности аргументов $\Delta t = t' - t$ (см. формулу 12.12); следовательно, при постоянном Δt корреляционная функция постоянна. Дисперсия стационарной функции не меняется со временем (см. формулу (12.11)), поэтому и нормированная корреляционная функция (см. формулу (12.14)) при постоянном Δt постоянна.

В табл. 12.7 постоянному Δt соответствует главная диагональ ($\Delta t = 0$) и параллели этой диагонали ($\Delta t = 50$, $\Delta t = 100$ и т. д.). Осредняя оценки значений нормированной корреляционной функции, приведенные в табл. 12.7, на главной диагонали ее параллелях, получаем оценку нормированной корреляционной функции (табл. 12.8).

Таблица 12.8

Δt	0	50	100	150	200	250	300	350
$r_x(\Delta t)$	1	0,79	0,76	0,73	0,71	0,66	0,56	0,51

Например, $r_x(50) = (0{,}82 + 0{,}60 + \ldots + 0{,}71)/7 = 0{,}79$; $r_x(100) = (0{,}87 + 0{,}68 + \ldots + 0{,}70)/6 = 0{,}76$ и т. д.

График функции $r_x(\Delta t)$ изображен на рис. 12.8 ломаной линией. При построении графика не следует забывать, что для стационарной случайной функции нормирования корреляционная функция — четная.

После определения достаточного числа значений эмпирической нормированной корреляционной функции $r(\Delta t)$ и построения графика обычно необходимо ее аппроксимировать (выравнить) простым аналитическим выражением. Аппроксимирующее выражение должно отображать наиболее характерные свойства графика функции и сглаживать случайные колебания при боль-

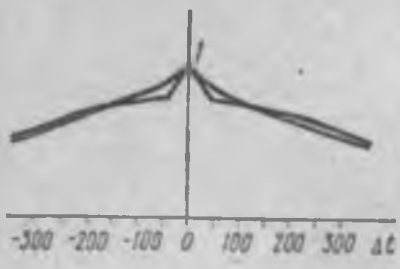

Рис. 12.8

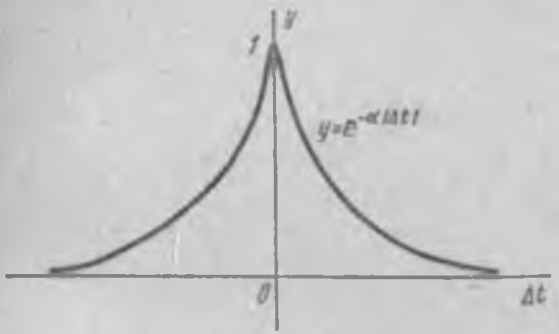

Рис. 12.9

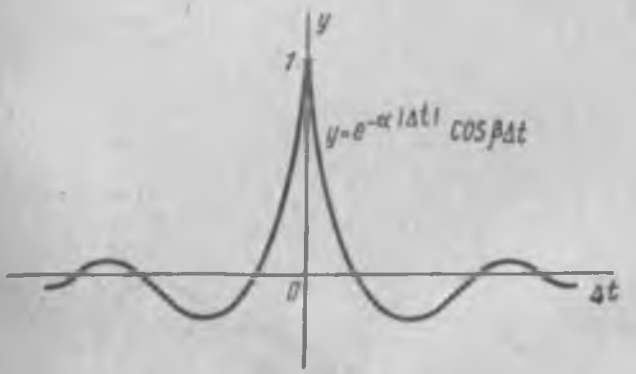

Рис. 12.10

ших значениях Δt, т. е. в точках, полученных осреднением небольшого числа данных и поэтому ненадежных.

Наиболее часто используют следующие аппроксимирующие выражения:

$$r(\Delta t) \approx e^{-\alpha|\Delta t|}, \qquad (12.15)$$

$$r(\Delta t) \approx e^{-\alpha|\Delta t|}\cos(\beta\Delta t). \qquad (12.16)$$

Графики функций $y=e^{-\alpha|\Delta t|}$ и $y=e^{-\alpha|\Delta t|}\cos(\beta\Delta t)$ приведены на рис. 12.9 и 12.10. Исследуемые в практических приложениях стационарные случайные функции, как правило, являются результатом суммирования большого числа независимых факторов, среди которых нет доминирующих. Для таких стационарных функций нормированные корреляционные функции имеют вид $\rho(\Delta t)=e^{-\alpha|\Delta t|}$ или $\rho(\Delta t)=e^{-\alpha|\Delta t|}\cos(\beta\Delta t)$.

Рассмотрим поведение функции $\rho(\Delta t)=e^{-\alpha|\Delta t|}$ при изменении α. При уменьшении α функция $\rho(\Delta t)$ убывает медленнее и, следовательно, характер изменения случайной функции более плавный. При увеличении α функция $\rho(\Delta t)$ убывает быстрее и характер колебания случайной функции более резкий и беспорядочный.

Поведение функции $\rho(\Delta t)=e^{-\alpha|\Delta t|}\cos(\beta\Delta t)$ зависит от соотношения параметров α и β, т. е. от того, что преобладает в корреляционной функции: убывание по закону $e^{-\alpha|\Delta t|}$ или колебание по закону $\cos(\beta\Delta t)$. Очевидно, при сравнительно малых α преобладает колебание, при сравнительно бо́льших — убывание.

Пример 12.4. Используя метод наименьших квадратов, выравнять оценку нормированной корреляционной функции профиля поверхности, заданную в табл. 12.8.

Р е ш е н и е. Величина ординаты профильной кривой колеблется в зависимости от пластических свойств обрабатываемой детали и инструмента, процесса резания, погрешностей измерения и других независимых друг от друга факторов, среди которых нет доминирующих. График функции $r_x(\Delta t)$, изображенный на рис. 12.8 ломаной линией, позволяет сделать вывод о том, что аппроксимирующее выражение имеет вид (12.15), т. е. $r_x(\Delta t)\approx e^{-\alpha|\Delta t|}$. Прологарифмировав обе части этого равенства, получим $\ln r_x(\Delta t)\approx -\alpha|\Delta t|$. В соответствии с методом наименьших квадратов коэффициент α определим из требования достижения минимума следующей функции:

$$f(\alpha)=\sum_{i=0}^{7}[-\alpha|\Delta t_i|-\ln r_x(\Delta t_i)]^2(8-i),$$

где $\Delta t_0=0$, $\Delta t_1=50,\ldots,\Delta t_7=350$, а множитель $(8-i)$ не что иное, как количество чисел, осреднением которых получено значение функции $r_x(\Delta t_i)$. Введение этого множителя позволяет при расчете α уменьшить роль значений $r_x(\Delta t_i)$, полученных осреднением небольшого числа данных и поэтому ненадежных. Действительно, для $r_x(\Delta t_1)=r_x(50)$, полученного осреднением семи чисел, множитель $(8-i)=(8-1)=7$; для $r_x(\Delta t_2)=r_x(100)$, полученного осреднением шести чисел, множитель $(8-i)=8-2=6$ и т. д.

Продифференцировав $f(\alpha)$ по α и приравняв производную к нулю, получим:

$$f'(\alpha) = \sum_{i=0}^{7} 2(8-i)[-\alpha|\Delta t_i| - \ln r_x(\Delta t_i)](-|\Delta t_i|) = 0.$$

Отсюда

$$\alpha = -\frac{\sum_{i=0}^{7}(8-i)|\Delta t_i| \ln r_x(\Delta t_i)}{\sum_{i=0}^{7}(8-i)|\Delta t_i|^2} =$$

$$= -\frac{8\cdot 0\cdot \ln r_x(0) + 7\cdot 50\cdot \ln r_x(50) + \ldots + 1\cdot 350\cdot \ln r_x(350)}{8\cdot 0^2 + 7\cdot 50^2 + \ldots + 1\cdot 350^2} =$$

$$= 0{,}002.$$

Значения функции $\rho(\Delta t) = e^{-0{,}002|\Delta t|}$ приведены в табл. 12.9. График ее изображен плавной линией на рис. 12.8.

Таблица 12.9

Δt	0	50	100	150	200	250	300	350
$e^{-0{,}002\|\Delta t\|}$	1	0,87	0,82	0,74	0,67	0,61	0,54	0,50

§ 12.6. Эргодические стационарные случайные функции

Рассмотренный в предыдущем параграфе метод определения характеристик стационарной случайной функции довольно сложен и состоит из двух этапов: оценки характеристик ординат случайной функции по реализациям и осреднения этих оценок. Возникает вопрос: нельзя ли для стационарной случайной функции этот двухступенчатый процесс заменить более простым, предположив, что единственная реализация достаточной продолжительности является достаточным опытным материалом для получения характеристик случайной функции? При более подробном рассмотрении этого вопроса оказывается, что не для всех стационарных функций одна реализация эквивалентна множеству реализаций. Рассмотрим, например, две стационарные случайные функции (см. рис. 12.1 и 12.11). Каждая реализация случайной функции, изображенной на рис. 12.1, обладает

одними и теми же характерными признаками: средним уровнем, вокруг которого происходят колебания, и средним размахом этих колебаний относительно среднего уровня. Очевидно, что одна из таких произвольно выбранных реализаций при достаточно большом времени испытания сможет дать нам достаточно хорошее представление о свойствах случайной функции. В частности,

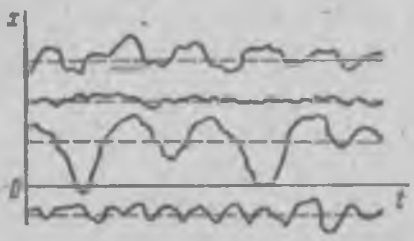

Рис. 12.11

осреднив значения этой реализации по времени, можно получить приближенное значение математического ожидания случайной функции. Про случайную функцию, каждая достаточно продолжительная реализация которой является как бы «полномочным представителем» всей совокупности возможных реализаций, говорят, что она обладает *эргодическим* свойством.

Рассмотрим случайную функцию, изображенную на рис. 12.11. Выберем произвольно одну реализацию, продолжим ее мысленно на достаточно длительный участок времени и осредним ее значения*. Очевидно, что средняя для каждой реализации может значительно отличаться от математического ожидания случайной функции. Про такую случайную функцию говорят, что она не обладает свойством эргодичности.

Если стационарная случайная функция $X(t)$, заданная на отрезке $[0, T]$, обладает эргодическим свойством, то:

---

* Если функция времени $f(t)$ задана на отрезке $[t_1, t_2]$, то осреднение ее значений по времени означает вычисление среднего значения функции на этом отрезке $\overline{f}(t)$ и производится так: $\overline{f}(t) = \frac{1}{t_2 - t_1} \times \int_{t_1}^{t_2} f(t)\,dt$, если $f(t)$ интегрируема.

1) ее математическое ожидание m_x приближенно равно средней по времени одной произвольно взятой реализации $x(t)$ достаточной продолжительности, т. е.

$$m_x \approx \frac{1}{T}\int_0^T x(t)\,dt; \qquad (12.17)$$

2) значение корреляционной функции $k_x(\Delta t)$ приближенно равно средней по времени из произведений отклонений ординат реализации $x(t)$ в точках, отстоящих друг от друга на величину Δt, от математического ожидания стационарной случайной функции m_x, т. е.

$$k_x(\Delta t) \approx \frac{1}{T-\Delta t}\int_0^{T-\Delta t} [x(t)-m_x][x(t+\Delta t)-m_x]\,dt. \qquad (12.18)$$

Здесь взято среднее значение по интервалу $[0, T-\Delta t]$, так как функции $x(t)$ и $x(t+\Delta t)$ известны вместе только в этом интервале. Формула (12.18) дает оценку корреляционной функции при $\Delta t>0$. При $\Delta t<0$ оценка корреляционной функции определяется из условия ее четности. Формулой (12.18) рекомендуется пользоваться при $\Delta t<T/5$. При бо́льших значениях Δt погрешность оценки корреляционной функции по формуле (12.18) увеличивается.

Если $\Delta t=0$, то $k_x(0)=D_x$ и дисперсия стационарной случайной функции, обладающей свойством эргодичности, может быть приближенно определена по формуле

$$D_x \approx \frac{1}{T}\int_0^T [x(t)-m_x]^2\,dt. \qquad (12.19)$$

Какие же стационарные случайные функции обладают, а какие не обладают эргодическим свойством? Оказывается, что неограниченное убывание корреляционной функции $k_x(\Delta t)$ стационарной случайной функции при $|\Delta t|\to\infty$ является достаточным условием для того, чтобы математическое ожидание функции можно было определять как среднее по времени, т. е. вычислять его по формуле (12.17). Если стационарная случайная функция к тому же нормально распределена, то неограниченное убывание корреляционной функции $k_x(\Delta t)$ при $|\Delta t|\to\infty$ является условием, достаточным для того, чтобы определить корреляционную функцию по формуле (12.18).

Нормально распределенные стационарные случайные функции с рассмотренными в предыдущем параграфе типовыми нормированными корреляционными функциями эргодичны, так как эти функции (а следовательно, и корреляционные функции) стремятся к нулю при $|\Delta t| \to \infty$. Поэтому их математические ожидания и корреляционные функции можно определить по одной достаточно продолжительной реализации по формулам (12.17), (12.18).

На практике, как правило, невозможно исследовать корреляционную функцию на бесконечном промежутке времени, а значения эмпирической корреляционной функции при больших Δt ненадежны, так как получаются осреднением небольшого числа данных. Кроме того, не надо забывать, что неограниченное убывание корреляционной функции является только достаточным условием эргодичности функции. Поэтому при решении практических задач вывод об эргодичности делается не столько на основании поведения корреляционной функции при $|\Delta t| \to \infty$, сколько на основании физических соображений, связанных с природой изучаемого процесса. Если процесс внутренне неоднороден и его можно разложить на более элементарные процессы, то он описывается только неэргодическими функциями. Наоборот, внутренняя однородность процесса говорит об его эргодичности. Например, процесс обработки поверхности детали при соблюдении одних и тех же основных условий от опыта к опыту описывается эргодической стационарной случайной функцией. Но если рассматривать процесс обработки поверхности в зависимости от скорости работы инструмента, т. е. от опыта к опыту менять скорость работы инструмента, то очевидно, что такой процесс уже не будет внутренне однородным, и хотя он может остаться стационарным, но свойством эргодичности обладать не будет.

§ 12.7. Определение характеристик эргодической стационарной случайной функции по одной реализации из опыта

Вычисляя характеристики эргодической стационарной случайной функции по одной реализации, используя формулы, приведенные в предыдущем параграфе, мы предполагали, что аргумент случайной функции t может

принимать любые значения на отрезке $[0, T]$ и что заранее известен аналитический вид реализации $x(t)$. При решении практических задач в результате опыта, как правило, оказываются известными значения реализации $x(t)$ для конечного множества значений аргумента и при вычислении характеристик эргодической стационарной случайной функции интегралы (12.17) и (12.18)

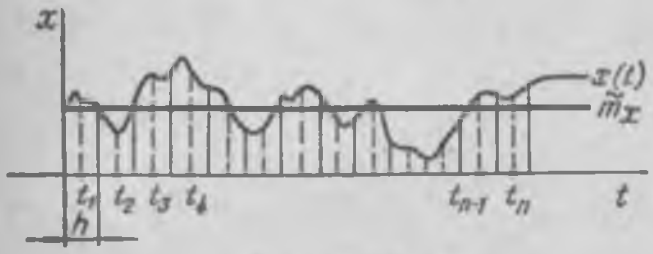

Рис. 12.12

заменяют конечными суммами. Покажем, как это делается.

Разобьем отрезок $[0, T]$ на n равных частей, длина которых $h=T/n$, и обозначим середины полученных участков через $t_1, t_2, \ldots, t_n$ (рис. 12.12). В формуле (12.17) для вычисления математического ожидания истолкуем интеграл как сумму площадей криволинейных трапеций с высотой h и средней линией, равной $x(t_i)$, где $i=1, 2, \ldots, n$. Тогда

$$m_x \approx \frac{1}{T}\int_0^T x(t)\,dt \approx \frac{1}{T}\sum_{i=1}^{n} hx(t_i) = \frac{1}{T}\sum_{i=1}^{n}\frac{T}{n}x(t_i),$$

$$m_x \approx \frac{1}{n}\sum_{i=1}^{n} x(t_i) = \bar{m}_x. \qquad (12.20)$$

Теперь покажем, как оценить значение корреляционной функции $k_x(\Delta t)$ при $\Delta t=0, h, \ldots, nh$. Пусть $\Delta t=mh$, где $m=0, 1, \ldots, n$. Тогда в соответствии с формулой (12.18)

$$k_x(mh) = \frac{1}{T-mh}\int_0^{T-mh} [x(t)-m_x][x(t+mh)-m_x]\,dt. \qquad (12.21)$$

Разделим интервал интегрирования, длина которого равна $T-mh=T-mT/n=(n-m)T/n$, на $n-m$ равных участков длиной $h=T/n$ и заменим в формуле (12.21) интеграл суммой площадей криволинейных трапеций. Получим

$$k_x(m, h) \approx \frac{1}{T-mh} \sum_{i=1}^{n-m} h\,[x(t_i)-m_x]\,[x(t_i+mh)-m_x].$$

Заменяя h величиной T/n, математическое ожидание m_x — его оценкой $\tilde{m}_x$ и учитывая, что $x(t_i+mh)=x(t_{i+m})$, найдем оценку корреляционной функции $k_x(\Delta t)$ в точке $\Delta t=mh=mT/n$:

$$k_x(mT/n) \approx \frac{1}{n-m} \sum_{i=1}^{n-m} [x(t_i)-\tilde{m}_x]\,[x(t_{i+m})-\tilde{m}_x] = k_x(mT/n). \qquad (12.22)$$

При $m=0$ получим оценку дисперсии эргодической стационарной случайной функции:

$$D_x \approx \frac{1}{n} \sum_{i=1}^{n} [x(t_i)-\tilde{m}_x]^2 = \tilde{D}_x. \qquad (12.23)$$

Для того чтобы, используя формулы (12.20) и (12.22), математическое ожидание и корреляционная функция эргодической стационарной случайной функции были определены с достаточной точностью, нужно, чтобы число точек n было достаточно велико (порядка сотни, а в некоторых случаях даже нескольких сотен). Количество точек определяется в зависимости от свойств случайной функции: для функции, изменяющейся сравнительно плавно, можно ограничиться меньшим числом точек, чем для функции, имеющей резкие и частые колебания.

§ 12.8. Об определении характеристик нестационарной случайной функции по одной реализации из опыта

Вычисление оценки математического ожидания эргодической стационарной случайной функции по формуле (12.20) можно рассматривать как сглаживание ее реализации, полученной в результате опыта (рис. 12.12).

Математическое ожидание нестационарной случайной функции можно также определить сглаживанием одной ее реализации (рис. 12.13), если полученная в результате опыта реализация случайной функции достаточной продолжительности и хорошо представляет всю совокупность возможных реализаций. Для сглаживания пользуются *методом скользящей средней*. Этот метод состоит

Рис. 12.13

в том, что за сглаженное значение функции в любой точке t принимают среднее значение в некотором интервале с центром в этой точке. При изменении t интервал скользит вдоль оси t, чем и объясняется название метода.

Разобьем отрезок $[0, T]$ записи реализации случайной функции на равные части точками $0=t_1, t_2, \ldots, t_k, \ldots, t_n = T$. Для определения сглаженного значения функции в точке t_k выделим из множества точек $t_1, t_2, \ldots, t_n$ такое подмножество $2p+1$ точек $t_{k-p}, \ldots, t_{k-1}, t_k, t_{k+1}, \ldots, t_{k+p}$, в котором центральной точкой является точка t_k. Сглаженное значение или, иначе говоря, *значение скользящей средней $(2p-1)$-го порядка в точке t_k* вычисляется по формуле

$$\bar{x}(t_k) = \frac{x(t_{k-p}) + \ldots + x(t_{k-1}) + x(t_k) + x(t_{k+1}) + \ldots + x(t_{k+p})}{2p+1} \qquad (12.24)$$

Используя формулу (12.24), можно вычислить сглаженные значения функции в точках $t_{p+1}, t_{p+2}, \ldots, t_{n-p}$.

Пример 12.5. Вычислить скользящую среднюю третьего порядка по данным подекадной реализации продукции (тыс. руб.), приведенным в табл. 12.10.

Решение. Скользящая средняя для II декады февраля равна (658+ +789+721)/3=723; для III декады— (789+721+488)/3=666 и т. д.

Как выбрать интервал сглаживания или, иначе говоря, каким должен быть порядок скользящей средней? Очевидно, чем больше интервал, тем сильнее сглаживание. Но, с другой стороны, при очень большом интервале сгладится и само математическое ожидание случайной функции. Поэтому интервал рекомендуется выбирать таким, чтобы при любом его расположении внутри отрезка $[0, T]$ реализация случайной функции имела на нем достаточно большое количество колебаний и в то же время чтобы математическое ожидание случайной функции на интервале можно было считать приблизительно линейным.

Оценку корреляционной функции нестационарной случайной функции по одной реализации можно получить, применяя метод скользящей средней к функции

$$z_{\Delta t}(t) = [x(t) - \bar{x}(t)][x(t + \Delta t) - \bar{x}(t + \Delta t)], \quad (12.25)$$

рассматриваемой при фиксированном Δt как функции от t. В формуле (12.25) функция $\bar{x}(t)$ и $\bar{x}(t+\Delta t)$ — сглаженная реализация соответственно на отрезке $[0, T-\Delta t]$ и $[\Delta t, T]$.

Таблица 12.10

Месяц	Февраль			Март			Апрель			Май			Июнь			Июль		
Декада	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
Реализация продукции, тыс. руб.	658	789	721	488	902	911	900	588	848	467	684	952	817	686	803	750	598	1046
Скользящая средняя		723	666	704	767	904	800	779	624	666	701	818	818	769	746	717	798	

Допустим, что сглаживание реализации $x(t)$ случайной функции проводилось скользящей средней $(2p+1)$-го порядка. Разобьем отрезок записи реализации случайной функции на равные части длиной h точками $0=t_1, t_2, \ldots, t_n=T$. Тогда оценкой значения корреляционной функции $k_x(\Delta t)$ в точке $\Delta t=mh$, где $m=0, 1, 2, \ldots, n$, является скользящая средняя, вычисленная по следующей совокупности чисел:

$$z_{\Delta t}(t_{p+1}) = [x(t_{p+1}) - \bar{x}(t_{p+1})][x(t_{p+1+m}) - \bar{x}(t_{p+1+m})],$$

$$z_{\Delta t}(t_{p+2}) = [x(t_{p+2} - \bar{x}(t_{p+2})][x(t_{p+2+m}) - \bar{x}(t_{p+2+m})],$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$z_{\Delta t}(t_{n-p-m}) = [x(t_{n-p-m}) - \bar{x}(t_{n-p-m})][x(t_{n-p}) - \bar{x}(t_{n-p})]. \quad (12.26)$$

§ 12.9. Цепи Маркова

До сих пор рассматривались особенности, связанные с изучением стационарных случайных функций. Другим не менее важным классом случайных функций являются *марковские случайные функции* — функции, дальнейшее поведение которых зависит только от значения, принятого функцией в настоящий момент, и не зависит от ранее принятых значений.

Рассмотрим только такие марковские случайные функции $X(t)$, у которых аргумент t принимает дискретное множество значений $t_0, t_1, t_2, \ldots, t_m$, и при каждом допустимом значении аргумента сама функция принимает одно из значений $x_0, x_1, \ldots, x_n$. Такие марковские случайные функции называют *цепями Маркова*. Очевидно, что для цепи Маркова вероятность того, что в момент t_i $(i=1, 2, \ldots, m)$ она примет значение $x_{q'}$ $(q'=1, 2, \ldots, n)$, зависит только от значения, принятого ею в предшествующий момент t_{i-1}. Обозначим вероятность того, что цепь Маркова в момент t_i примет значение $x_{q'}$ при условии, что в момент t_{i-1} она приняла значение x_q, через $p_{qq'}(t_{i-1}, t_i)$. Вероятность $p_{qq'}(t_{i-1}, t_i)$ называется *вероятностью перехода из состояния xx_q в состояние $x_{q'}$ за интервал* (t_{i-1}, t_i). В дальнейшем будем рассматривать цепи Маркова с равноудаленными значениями аргумента $t_0, t_1, t_2, \ldots, t_m$.

Цепь Маркова называется *однородной по времени*, если вероятности перехода $p_{qq'}(t_{i-1}, t_i)$ из состояния x_q

в состояние $x_{q'}$ за промежуток времени от t_{i-1} до t_i, где $q, q'=0, 1, \ldots, n$; $i=1, 2, \ldots, m$, зависят от длины промежутка $\tau=t_i-t_{i-1}$ и не зависят от начала отсчета t_i, т. е. $p_{qq'}(t_{i-1}, t_i)=p_{qq'}(\tau)$.

Вероятности перехода за интервал времени τ можно свести в матрицу $\Gamma(\tau)$:

$$\Gamma(\tau)=\begin{vmatrix} p_{00} & p_{01} & \cdots & p_{0n} \\ p_{10} & p_{11} & \cdots & p_{1n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ p_{n0} & p_{n1} & & p_{nn} \end{vmatrix}.$$

Эту матрицу называют *матрицей перехода*. Заметим, что элементы матрицы неотрицательны, а сумма элементов любой строки, например q-й, равна единице, т. е.

$$\sum_{q'=0}^{n} p_{qq'}=1, \tag{12.27}$$

так как за интервал времени τ цепь Маркова из состояния x_q обязательно перейдет в одно из допустимых состояний $x_0, x_1, \ldots, x_n$.

Обозначим через $p_q(t_i)$ вероятность того, что $X(t)$ в момент t_i принимает значение x_q. Вектор $p(t_i)=[p_0(t_i), p_1(t_i), \ldots, p_n(t_i)]$ называют *вектором вероятностей* состояний цепи Маркова в момент t_i. Очевидно, что элементы вектора неотрицательны, а сумма элементов равна единице, т. е.

$$\sum_{q=0}^{n} p_q(t_i)=1, \tag{12.28}$$

так как в момент время t_i цепь Маркова $X(t)$ обязательно принимает одно из значений $x_0, x_1, \ldots, x_n$.

Теорема 12.1. *Вектор вероятностей состояний цепи Маркова в момент t_i равен произведению вектора вероятностей состояний в момент t_{i-1} на матрицу перехода, т. е.*

$$\mathrm{p}(t_i)=\mathrm{p}(t_{i-1})\Gamma(\tau). \tag{12.29}$$

Доказательство Докажем справедливость теоремы для цепи Маркова с двумя допустимыми состоя-

ниями x_0 и x_1. При этих условиях $\Gamma(\tau) = \begin{vmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{vmatrix}$. Докажем, что

$$\mathrm{p}(t_i) = \mathrm{p}(t_{i-1}) \cdot \Gamma(\tau). \tag{12.30}$$

Найдем вероятность $p_0(t_i)$ того, что в момент времени t_i цепь принимает значение x_0. К этому могут привести два события: событие A, состоящее в том, что в момент t_{i-1} цепь находилась в состоянии x_0 и за время τ не вышла из него; событие B, состоящее в том, что в момент t_{i-1} цепь находилась в состоянии x_1 и за время τ перешла в состояние x_0. По теореме сложения вероятностей несовместных событий,

$$p_0(t_i) = P(A) + P(B).$$

Вероятность события A найдем по теореме умножения вероятностей. Вероятность того, что в момент t_{i-1} цепь принимает значение x_0, равна $p_0(t_{i-1})$. Вероятность того, что она за время τ не изменит значения x_0, в силу свойства марковости обусловлена только ее значением в предшествующий момент t_{i-1} и равна p_{00}. Следовательно, $P(A) = p_0(t_{i-1})\, p_{00}$. Вероятность события B также найдем по теореме умножения вероятностей $P(B) = p_1(t_{i-1})\, p_{10}$.

Следовательно,

$$p_0(t_i) = p_0(t_{i-1})\, p_{00} + p_1(t_{i-1})\, p_{10}. \tag{12.31}$$

Рассуждая аналогично, можно показать, что

$$p_1(t_i) = p_0(t_{i-1})\, p_{01} + p_1(t_{i-1})\, p_{11}. \tag{12.32}$$

Таким образом, вектор вероятностей состояний цепи Маркова в момент t_i

$$\mathrm{p}(t_i) = [p_0(t_i),\, p_1(t_i)] = [p_0(t_{i-1})\, p_{00} + p_1(t_{i-1})\, p_{10};$$
$$p_0(t_{i-1})\, p_{01} + p_1(t_{i-1})\, p_{11}].$$

С другой стороны, произведение

$$\mathrm{p}(t_{i-1}) \cdot \Gamma(\tau) = [p_0(t_{i-1});\, p_1(t_{i-1})] \cdot \begin{vmatrix} p_{00}, & p_{01} \\ p_{10}, & p_{11} \end{vmatrix} =$$
$$= [p_0(t_{i-1})\, p_{00} + p_1(t_{i-1})\, p_{10};\; p_0(t_{i-1})\, p_{01} + p_1(t_{i-1})\, p_{11}] =$$
$$= [p_0(t_i);\; p_1(t_i)] = \mathrm{p}(t_i). \quad \square$$

□ Доказательство теоремы для более общего случая не представляет труда.

Следствие. *Однородная цепь Маркова определяется вектором вероятностей* $\mathbf{p}(t_0)$ *состояний в начальный момент и матрицей перехода* $\Gamma(\tau)$.

Доказательство. Матрица перехода $\Gamma(\tau)$ для однородной цепи Маркова зависит от длины промежутка $\tau=t_i-t_{i-1}$ и не зависит от начала отсчета t_i. Так как значения $t_0, t_1, \ldots, t_m$, аргумента равноудалены, то равенство (12.29) справедливо для любого $i=1, 2, \ldots, m$ и является рекуррентным. Следовательно, справедлива следующая цепочка равенств:

$$\mathbf{p}(t_i) = \mathbf{p}(t_{i-1})\,\Gamma(\tau) = \mathbf{p}(t_{i-2})\,\Gamma^2(\tau) = \\ = \mathbf{p}(t_{i-3})\,\Gamma^3(\tau) = \cdots = \mathbf{p}(t_0)\,\Gamma^i(\tau),$$

где $\Gamma^i(\tau)$ — i-я степень матрицы перехода. Таким образом,

$$\mathbf{p}(t_i) = \mathbf{p}(t_0)\,\Gamma^i(\tau). \qquad (12.33)$$

Из равенства (12.33) следует, что вектор вероятностей $\mathbf{p}(t_i)$, иначе говоря, закон распределения ординаты $X(t_i)$, для любого момента времени $t_i (i=1, 2, \ldots, m)$ полностью определяется вектором $\mathbf{p}(t_0)$ и матрицей $\Gamma(\tau)$.

Матрица $\Gamma^i(\tau)$ является матрицей вероятностей перехода $p_{qq'}$ $(i\tau)$ за интервал времени равный $i\tau$. Она обладает теми же свойствами, что и матрица $\Gamma(\tau)$: элементы ее неотрицательны и сумма элементов каждой строки равна единице.

Пример 12.6. Пусть имеются три конкурирующих изделия x_0, x_1 x_2. Для определения спроса на эти изделия произведен в некоторый момент t_0 опрос 1000 человек. Оказалось, что изделие x_0 покупают 500 человек, изделие x_1 — 200, а x_2 — 300 человек. Если через $p_q(t_0)$ обозначить статистическую вероятность потребности в изделии x_q в момент t_0, то $\mathbf{p}(t_0) = (0{,}5;\ 0{,}2;\ 0{,}3)$.

Предположим, что поведение покупателей в каждый следующий месяц обусловлено только их поведением в предыдущий месяц (таким образом, исследуется цепь Маркова). По истечении месяца оказалось, что из 500 человек, покупавших изделие x_0, 450 человек продолжают его покупать, 40 человек стали покупать изделие x_1 и 10 — изделие x_2, т. е. статистические вероятности $p_{00}=450/500=0{,}9$; $p_{01}=40/500=0{,}08$; $p_{02}=10/500=0{,}02$.

Из 200 человек, покупавших изделие x_1, 60 человек продолжают его покупать, 80 стали покупать изделие x_0, 60 — изделие x_2, т. е. $p_{10}=0{,}4$; $p_{11}=0{,}3$; $p_{12}=0{,}3$.

Из 300 человек, покупавших изделие x_2, 60 человек продолжают его покупать, 210 человек стали покупать изделие x_0, 30 — изделие x_1, т. е. $p_{20}=0{,}7$; $p_{21}=0{,}1$; $p_{22}=0{,}2$.

Определить, какое изделие пользуется по истечении месяца наибольшим спросом.

Решение. В соответствии с формулой (12.29) имеем

$$p(t_1) = p(t_0)\Gamma(\tau) = (0{,}5;\ 0{,}2;\ 0{,}3)\begin{vmatrix} 0{,}9 & 0{,}08 & 0{,}02 \\ 0{,}4 & 0{,}3 & 0{,}3 \\ 0{,}7 & 0{,}1 & 0{,}2 \end{vmatrix} =$$

$$= (0{,}74;\ 0{,}13;\ 0{,}13).$$

Таким образом, через месяц наибольшим спросом будет пользоваться изделие x_0. Если предположить, что поведение покупателей со временем не меняется, т. е. что цепь однородна по времени, то по формуле (12.33) можно определить, какое изделие будет пользоваться наибольшим спросом по истечении двух, трех месяцев и т. д.

§ 12.10. Эргодическое свойство однородных цепей Маркова

Рассмотрим предел вероятности перехода $p_{qq'}(i\tau)$ однородной цепи Маркова из состояния x_q в состояние $x_{q'}$ за интервал времени $i\tau$ при $i \to \infty$. Вполне естественно предположить, что по мере увеличения интервала $i\tau$ влияние начального состояния x_q уменьшается и при достаточно большом интервале станет ничтожно малым, т. е.

$$\lim_{i\to\infty} p_{qq'}(i\tau) = p_{q'}. \qquad (12.34)$$

Однородная цепь Маркова, для любых двух состояний которой справедливо равенство (12.34), называется *эргодической*. Марковым было сформулировано *достаточное условие эргодичности однородных цепей с конечным числом состояний: если существует такое* $\tau > 0$, *что при любых допустимых* q *и* q' *вероятности* $p_{qq'}$ *положительны, то цепь эргодична.* Учитывая определение эргодичности (12.34), имеем

$$\lim_{i\to\infty} \Gamma^i(\tau) = \lim_{i\to\infty} \Gamma(i\tau) = \lim_{i\to\infty} \begin{vmatrix} p_{00}(i\tau), \dots, p_{0n}(i\tau) \\ p_{10}(i\tau), \dots, p_{1n}(i\tau) \\ \dots\dots\dots\dots \\ p_{n0}(i\tau), \dots, p_{nn}(i\tau) \end{vmatrix} =$$

$$= \begin{vmatrix} p_0, p_1, \dots, p_n \\ p_0, p_1, \dots, p_n \\ \dots\dots\dots \\ p_0, p_1, \dots, p_n \end{vmatrix},$$

т. е. если цепь эргодична, то i-я степень матрицы перехода при $i \to \infty$ стремится к предельной матрице с одинаковыми строками. Можно показать, что все элементы предельной матрицы положительны и что

$$\sum_{q'=0}^{n} p_{q'} = 1. \tag{12.35}$$

Переходя к пределам в обеих частях равенства (12.33) и учитывая равенство (12.28), получаем

$$\lim_{i\to\infty} \mathbf{p}(t_i) = \lim_{i\to\infty} [\mathbf{p}(t_0)\,\mathbf{\Gamma}^i(\tau)] = \mathbf{p}(t_0) \lim_{i\to\infty} \mathbf{\Gamma}^i(\tau) =$$

$$= [p_0(t_0),\ p_1(t_0), \ldots, p_n(t_0)] \begin{vmatrix} p_0, & p_1, \ldots, & p_n \\ p_0, & p_1, \ldots, & p_n \\ \cdots & \cdots & \cdots \\ p_0, & p_1, \ldots, & p_n \end{vmatrix} =$$

$$= \left[p_0 \sum_{q=0}^{n} p_q(t_0);\ p_1 \sum_{q=0}^{n} p_q(t_0); \ldots; p_n \sum_{q=0}^{n} p_q(t_0) \right] =$$

$$= (p_0, p_1, \ldots, p_n).$$

Таким образом,

$$\lim_{i\to\infty} \mathbf{p}(t_i) = (p_0,\ p_1, \ldots, p_n). \tag{12.36}$$

Переходя к пределам в обеих частях равенства (12.29), получаем

$$\lim_{i\to\infty} \mathbf{p}(t_i) = \lim_{i\to\infty} [\mathbf{p}(t_{i-1})\,\mathbf{\Gamma}(\tau)].$$

Учитывая равенство (12.36), имеем

$$(p_0, p_1, \ldots, p_n) = (p_0, p_1, \ldots, p_n)\,\mathbf{\Gamma}(\tau). \tag{12.37}$$

Допустим, что матрица перехода $\Gamma(\tau)$ известна. Выясним, нельзя ли, используя соотношение (12.37), определить вектор предельных вероятностей $(p_0, p_1, \ldots, p_n)$? Запишем равенство (12.37) в следующем виде:

$$\begin{cases} p_0 = p_0 p_{00} + p_1 p_{10} + \cdots + p_n p_{n0}, \\ p_1 = p_0 p_{01} + p_1 p_{11} + \cdots + p_n p_{n1} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ p_n = p_0 p_{0n} + p_1 p_{1n} + \cdots + p_n p_{nn}. \end{cases} \tag{12.38}$$

Учитывая соотношение (12.27), можно показать, что система уравнений (12.38) линейно зависима. Заменив

любое из уравнений системы (12.38) уравнением (12.35), получим систему $n+1$ линейно независимых уравнений с неизвестными $p_0, p_1, \ldots, p_n$. В результате ее решения определим предельные вероятности

Пример 12.7. В условиях примера 12.6 определить, какое изделие будет пользоваться наибольшим спросом по истечении достаточно продолжительного периода.

Р е ш е н и е. Все элементы матрицы перехода в примере 12.7 положительны, т. е. условие эргодичности выполняется, следовательно, предельные вероятности p_0, p_1, p_2 потребления изделий существуют. Система уравнений (12.38) в данном случае имеет вид

$$\begin{cases} p_0 = p_0 \cdot 0{,}9 + p_1 \cdot 0{,}4 + p_2 \cdot 0{,}7, \\ p_1 = p_0 \cdot 0{,}08 + p_1 \cdot 0{,}3 + p_2 \cdot 0{,}1, \\ p_2 = p_0 \cdot 0{,}02 + p_1 \cdot 0{,}3 + p_2 \cdot 0{,}2. \end{cases}$$

Система линейно зависима. Заменяя третье уравнение системы уравнением $p_0+p_1+p_2=1$, получаем систему

$$\begin{cases} p_0 = p_0 \cdot 0{,}9 + p_1 \cdot 0{,}4 + p_2 \cdot 0{,}7, \\ p_1 = p_0 \cdot 0{,}08 + p_1 \cdot 0{,}3 + p_2 \cdot 0{,}1, \\ p_0 + p_1 + p_2 = 1. \end{cases}$$

Ее решения: $p_0=0{,}84$; $p_1=0{,}10$, $p_2=0{,}06$. Итак, по истечении достаточно продолжительного периода наибольшим спросом будет пользоваться изделие x_0.

В заключение заметим, что формулы, полученные для цепей Маркова, можно обобщить и на другие разновидности марковских случайных функций. Так, при изучении марковских функций с непрерывным временем вводят прерывное время меняющееся скачкообразно через интервал Δt, а затем переходят к пределу при $\Delta t \to 0$. Формулы цепей Маркова также можно обобщить и на случай непрерывного изменения значений случайной функции.

Глава 13

Моделирование случайных функций на ЭВМ

§ 13.1. Общая идея метода статистического моделирования

В терминах случайных функций формулируется большое число научных и практических задач (многие задачи нейтронной физики, исследования различных систем управления со случайными воздействиями на входах и т. д.). Решить такие задачи аналитическими методами и связать заданные условия с исходом удается далеко не всегда.

В тех случаях, когда найти аналитическое решение трудно, используют метод статистического моделирования, сводящийся к построению искусственных реализаций случайных функций. Метод статистического моделирования называют также методом статистических испытаний (так как построение реализаций случайной функции — это, по существу, проведение статистического эксперимента, испытания) или методом Монте-Карло.

Воспроизвести реальные процессы, протекающие в той или иной системе, можно, используя различные средства: эксперименты с подбрасыванием монеты и игральной кости, таблицы случайных чисел с различными законами распределения и т. д. Однако только с появлением ЭВМ метод статистических испытаний смог найти широкое применение, поскольку построение искусственных реализаций случайных процессов — очень трудоемкая работа. Чтобы решить задачи методом Монте-Карло, нужно иметь источник получения чисел, соответствующих случайным величинам с различными законами распределения.

Оказывается, что основную роль играют случайные величины, распределенные по равномерному закону на интервале (0, 1). Условимся для краткости случайную величину, равномерно распределенную на интервале (0, 1), называть случайным числом и обозначать R.

С помощью случайных чисел можно моделировать результаты испытаний в схеме случайных событий, получать значения случайных величин с заданным законом распределения, формировать реализации случайных функций.

§ 13.2. Получение случайных чисел на интервале (0, 1)

Существует три способа получения таких чисел: таблицы случайных чисел, генераторы случайных чисел и метод псевдослучайных чисел.

Таблицей случайных чисел называют таблицу, содержащую чередующиеся в случайном порядке цифры 0, 1, 2,..., 9. Случайность порядка этих цифр должна гарантироваться методом составления таблиц. Именно: каждая из цифр должна иметь одинаковую возможность попасть в таблицу, и отбор каждой следующей цифры не должен зависеть от результатов предыдущих отборов. Таким методом является метод случайной выборки с возвратом из 10 цифр по одной цифре. Полученные в результате отбора цифры записывают в виде таблицы (см. табл. 4 приложений, где для удобства цифры объединены в группы по 5 шт.).

Таблицу случайных чисел можно ввести в запоминающее устройство ЭВМ и, если в ходе расчета понадобится значение случайной величины X с распределением

X	0	1	...	9
P	0,1	0,1	...	0,1

то берут очередную цифру из этой таблицы. Если понадобится значение случайной величины с распределением

X	0	1	...	99
P	0,01	0,01	...	0,01

то берут очередную пару цифр, и т. д.

Если ЭВМ работает с m разрядными десятичными числами, то, разделив очередную совокупность m цифр, взятую из таблицы случайных чисел, на 10^m, получим достаточно хорошее приближение к значению случайной величины, равномерно распределенной на интервале (0, 1).

Таблицы случайных чисел используют только при расчетах по методу Монте-Карло вручную (все ЭВМ обладают сравнительно малой внутренней памятью и постоянное обращение к ней замедляет счет). В качестве генератора случайных чисел на ЭВМ чаще всего используются шумы в электронных лампах: если за некоторый промежуток времени уровень шума превысил заданный порог четное число раз, то в двоичный разряд специальной ячейки записывается нуль, а если нечетное число раз, то единица. При параллельной работе m генераторов в ячейке получается одно m-разрядное двоичное число, которое дает достаточно хорошее приближение к значению равномерно распределенной случайной величины.

Однако и этот способ не лишен недостатков. Во-первых, в результате неисправностей нули и единицы в каком-либо из разрядов могут появляться не одинаково часто и вырабатываемую последовательность чисел уже нельзя считать случайной, во-вторых, воспроизвести полученную последовательность невозможно, что необходимо при двойном просчете задачи с целью исключения влияния случайных сбоев.

Для получения случайных чисел на интервале (0, 1) используют также метод псевдослучайных чисел. Псевдослучайными называют числа, вычисляемые по некоторому алгоритму и обладающие следующей особенностью (позволяющей использовать их в качестве случайных чисел): частотные свойства последовательности псевдослучайных чисел в достаточной степени близки к свойствам равномерно распределенных на интервале (0, 1) случайных последовательностей.

Первый алгоритм для получения псевдослучайных чисел на ЭВМ был предложен Дж. Нейманом. Алгоритм

состоит в том, что берут некоторое произвольное число и возводят в квадрат, в полученном результате отбрасывают цифры с обоих концов и этот процесс возведения в квадрат и отбрасывания цифр повторяют. Например,

пусть $R_0 = 0{,}2061$, тогда $R_0^2 = 0{,}04\,|\,2477\,|\,21$;

» $R_1 = 0{,}2477$, » $R_1^2 = 0{,}06\,|\,1355\,|\,29$;

» $R_2 = 0{,}1355$, » $R_2^2 = 0{,}01\,|\,8360\,|\,25$ и т. д.

Существуют и другие алгоритмы получения псевдослучайных чисел. Достоинства метода псевдослучайных чисел очевидны. Во-первых, на получение каждого числа затрачивается всего несколько простых операций, поэтому скорость нахождения псевдослучайных чисел имеет тот же порядок, что и скорость ЭВМ; во-вторых, программа алгоритма занимает всего несколько ячеек памяти; в-третьих, полученную последовательность чисел легко воспроизвести; в-четвертых, соответствие частотных свойств вырабатываемой последовательности чисел свойствам равномерно распределенных на интервале (0, 1) случайных чисел достаточно проверить лишь один раз, затем эту последовательность можно использовать много раз.

§ 13.3. Моделирование дискретной случайной величины

Допустим, что требуется получать значения случайной величины X со следующим распределением:

X	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

Разобьем интервал (0, 1) на n интервалов, длины которых $p_1, p_2, ..., p_n$ (рис. 13.1). Если случайное число R попадает на интервал p_i, то считают, что случайная величина X принимает значение x_i.

Для обоснования правомерности такой процедуры убедимся, что вероятность попадания R в интервал p_i, т. е. вероятность события $\sum_{q=1}^{i-1} p_q < R < \sum_{q=1}^{i} p_q$, равна $P(X = x_i) = p_i$. Здесь R — случайная величина, равномерно распределенная на интервале (0, 1), поэтому ее функ-

Далее, учитывая, что функция распределения равномерной на интервале (0, 1) случайной величины имеет вид $F_R(r)=r$, если $0\leqslant r\leqslant 1$, запишем следующую цепочку равенств: $P(X<x)=P(R<F(x))=F_R[F(x)]=F(x)$. Итак, если X — корень уравнения (13.1), то $P(X<x)=F(x)$. □

Таким образом, определение значения случайной величины X с функцией распределения $F(x)$ сводится к следующей процедуре: получаем случайное число R на интервале (0, 1) и вычисляем $F^{-1}(R)$, где F^{-1} — функция, обратная к функции F.

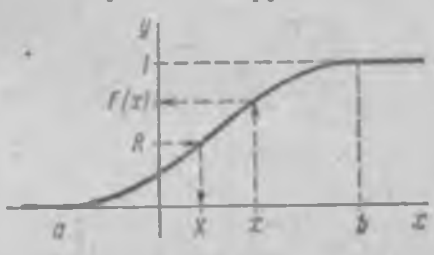

Рис. 13.4

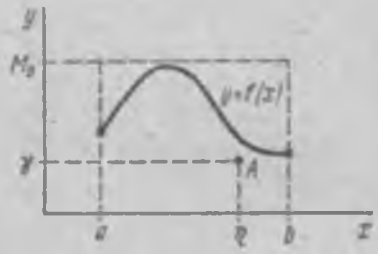

Рис. 13.5

Пример 13.2. Определить последовательность значений случайной величины X, распределенной по показательному закону.

Решение. Для случайной величины X функция распределения $F(x)=1-e^{-\lambda x}$ $(x>0)$. Корнем уравнения $1-e^{-\lambda x}=R$ является $X=(-1/\lambda)\ln(1-R)$. Так как R — случайная величина с равномерным распределением на интервале (0, 1), то и $(1-R)$ также случайная величина с равномерным распределением на интервале (0, 1), поэтому можно взять $X=(-1/\lambda)\ln R$.

Последовательности случайных чисел R_1, R_2, ... соответствует последовательность $(-1/\lambda)\ln R_1$, $(-1/\lambda)\ln R_2$, ... значений случайной величины X с показательным законом распределения.

Может оказаться, что решение уравнения (13.1) относительно X трудно. В этой ситуации для моделирования непрерывной случайной величины может быть использован метод Неймана. Предположим, что случайная величина X определена на интервале (a, b) и плотность ее распределения ограничена: $f(x)\leqslant M_0$ (рис. 13.5). Метод Неймана сводится к следующей процедуре:

1) выбирают два случайных числа R_1 и R_2 на интервале (0, 1) и строят точку $A(\eta;\ \gamma)$ с координатами $\eta=a+R_1(b-a)$, $\gamma=R_2M_0$;

2) если точка A лежит под кривой $y=f(x)$, т. е. $f(\eta)\geqslant\gamma$, то полагают $x=\eta$; если же точка A лежит над кривой $y=f(x)$, т. е. $f(\eta)<\gamma$, то точку $A(\eta,\ \gamma)$ отбрасывают и выбирают новую пару случайных чисел.

Эффективность метода Неймана характеризуется величиной отношения площади криволинейной трапеции под кривой $y=f(x)$, равной $\int_a^b f(x)\,dx$, к площади всего прямоугольника, равной $(b-a)M_0$. Если это отношение мало, то значительная часть машинного времени будет расходоваться на получение точек, лежащих над кривой и отбрасываемых в процессе вычислений. Отсюда следует, в частности, что для экспоненциального распределения этот метод не эффективен, наибольшую же эффективность метод имеет для равномерного распределения. На практике метод Неймана обычно используют в том случае, когда указанное отношение площадей превышает величину 0,3.

Помимо рассмотренных общих методов моделирования непрерывных случайных величин (решение уравнения (13.1) и метод Неймана) для многих законов распределения существуют специальные алгоритмы, работающие более быстро по сравнению с общими методами.

§ 13.5. Моделирование нормально распределенной случайной величины

Допустим, что нужно получать значения нормально распределенной случайной величины X, математическое ожидание которой $M(X)=a$, а дисперсия $\sigma_x^2=\sigma^2$.

Рассмотрим метод моделирования нормально распределенной случайной величины, основанный на центральной предельной теореме. Согласно этой теореме, при сложении достаточно большого числа независимых случайных величин получается случайная величина, распределенная приближенно по нормальному закону. Опыт показывает, что при сложении всего шести случайных величин, равномерно распределенных на интервале (0, 1), получается случайная величина, которая с точностью, достаточной для большинства прикладных задач, может считаться нормальной.

Отсюда возникает такой метод моделирования нормально распределенной случайной величины X: сложить шесть случайных чисел, пронормировать эту сумму, т. е. получить случайную величину T с $M(T)=0$ и $\sigma_t^2=1$, тогда $X=\sigma T+a$.

ция распределения $F_R(r)=r$, если $0\leqslant r\leqslant 1$. Учитывая это, запишем следующую цепочку равенств: $P\left(\sum_{q=1}^{i-1} p_q < R < \sum_{q=1}^{i} p_q\right) = F_R\left(\sum_{q=1}^{i} p_q\right) - F_R\left(\sum_{q=1}^{i-1} p_q\right) = p_i$. □

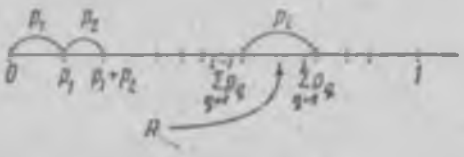

Рис. 13.1

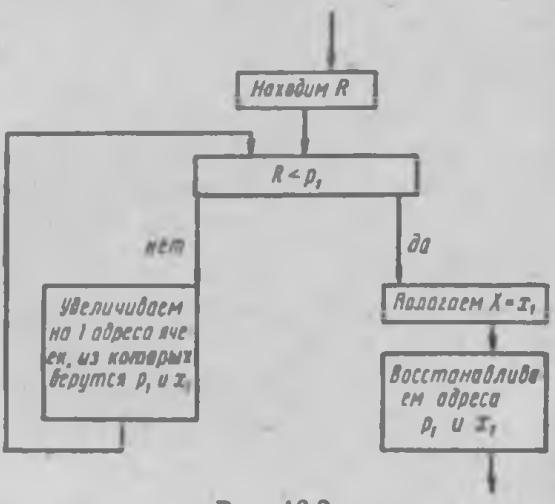

Рис. 13.2

Блок-схема программы получения значения дискретной случайной величины X приведена на рис. 13.2. Предполагается, что числа $x_1, x_2,..., x_n$ и вероятности $p_1, p_1+p_2, p_1+p_2+p_3,..., 1$ располагаются в ячейках памяти подряд. Последовательности случайных чисел $R_1, R_2,...$ соответствует последовательность значений случайной величины X с заданным выше законом распределения.

Используя процедуру моделирования дискретной случайной величины, можно моделировать результат испытания в схеме случайных событий, т. е. ответить на вопрос: какое из несовместных событий $A_1,..., A_n$, образующих полную группу, произойдет, если $P(A_i)=p_i (i=1, 2,..., n)$.

Пример 13.1. В результате испытания может произойти одно из событий A_1, A_2, A_3. Проводится серия независимых испытаний. Определить, какие события появятся, если $P(A_1)=0{,}35$, $P(A_2)=0{,}25$, $P(A_3)=0{,}4$.

Решение. Разделим интервал (0,1) на три интервала, как показано на рис. 13.3. Возьмем из табл. 4 приложений любое число. Пусть $R=0{,}100$; это число попало на первый интервал — произошло событие A_1. Следующее случайное число $R=0{,}325$ — опять произошло событие A_1: далее $R=0{,}765$ — произошло событие A_3 и т. д.

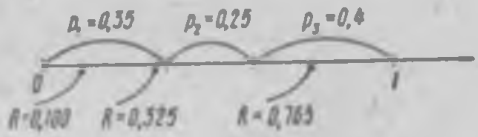

Рис. 13.3

Рассмотренная процедура моделирования дискретной случайной величины, вообще говоря, может быть использована для нахождения значений дискретной величины с любым законом распределения. Однако для целого ряда законов распределения, учитывая их специфику, можно построить более быстрые и легче программируемые процедуры моделирования.

§ 13.4. Моделирование непрерывной случайной величины

Допустим, что нужно получать значения непрерывной случайной величины X, определенной на интервале (a, b), если известна ее функция распределения $F(x)$. Убедимся в том, что случайная величина X является решением уравнения

$$F(X)=R. \qquad (13.1)$$

Дадим графическое толкование решения этого уравнения (рис. 13.4) и убедимся в том, что оно единственное. Из свойств функции $F(x)$ известно, что если случайная величина X задана на интервале (a, b), то $F(x)$ — монотонно возрастающая функция на этом интервале, поэтому любая прямая $y=R$ пересекает график $y=F(x)$ в единственной точке, абсцисса которой X и является решением уравнения (13.1).

Пусть X — корень уравнения (13.1); найдем его функцию распределения, т. е. вероятность $P(X<x)$. Из рис. 13.4 видно, что если $X<x$, то $R<F(x)$; верно и обратное утверждение. Значит, $P(X<x)=P(R<F(x))$.

Пусть $Z=\sum_{i=1}^{6} R_i$, где $R_1, R_2, \ldots, R_6$ — независимые равномерно распределенные на интервале (0, 1) случайные величины. Учитывая, что $M(R_i)=1/2$ и $\sigma^2_{R_i}=1/12$, получаем $M_Z=\sum_{i=1}^{6} M(R_i)=3$, а $\sigma_Z^2=\sum_{i=1}^{6}\sigma^2_{R_i}=1/2$. Пронормируем Z, т. е. перейдем от нее к величине $T=(Z-M(z))/\sigma_z=(Z-3)\sqrt{2}$, а затем найдем

$$X=\sigma(Z-3)\sqrt{2}+a=\sqrt{2}\,\sigma\left(\sum_{i=1}^{6} R_i-3\right)+a.$$

Таким образом, чтобы определить значение нормальной случайной величины X с математическим ожиданием a и дисперсией σ^2, нужно взять шесть случайных чисел, сложить их, из суммы вычесть 3, результат умножить на $\sigma\sqrt{2}$ и прибавить a.

§ 13.6. Моделирование системы случайных величин и случайных функций

Допустим, что требуется определить значения непрерывных случайных величин X и Y, заданных совместной функцией плотности $f(x, y)$. Если случайные величины X и Y независимы, то, как было показано в § 3.9, совместная функция плотности $f(x, y)=f(x)f(y)$, где $f(x)$ и $f(y)$ — функции плотности случайных величин X и Y, и моделирование системы случайных величин X и Y просто сводится к моделированию каждой случайной величины в отдельности. Если случайные величины X и Y зависимы, то $f(x,y)=f(x)f(y|x)$, где $f(y|x)=f(x,y)/f(x)$ — условная функция плотности Y при $X=x$, и моделирование системы случайных величин сводится к следующей процедуре:

1) определяют значение случайной величины X с функцией плотности $f(x)$;

2) найденное значение величины X подставляют в выражение для условной функции плотности $f(y|x)$;

3) определяют значение случайной величины, функция плотности которой равна $f(y|x)$, это и есть значение случайной величины Y.

Если случайные величины X и Y дискретны, то процедура их моделирования аналогична рассмотренной с той лишь разницей, что вместо функций плотности рассматриваются ряды распределения и используются алгоритмы моделирования дискретных случайных величин.

Для моделирования системы непрерывных случайных величин может быть также использован *метод Неймана*. Пусть система случайных величин $X_1, X_2, ..., X_n$ имеет ограниченную функцию плотности $f(x_1, x_2, ..., x_n) \leqslant M_0$ и $a_i < x_i < b_i$, где $i=1, 2, ..., n$. Метод Неймана в этом случае сводится к следующей процедуре:

1) берут $(n+1)$-е случайное на интервале (0, 1) число $R_1, R_2, ..., R_{n+1}$ и рассматривают величины $\eta_i = a_i + R_i(b_i - a_i)$, где $i=1, 2, ..., n$, и величину $\gamma = M_0 R_{n+1}$;

2) если $f(\eta_1, \eta_2, ..., \eta_n) \geqslant \gamma$, то полагают $X_1 = \eta_1$, $X_2 = \eta_2$, ..., $X_n = \eta_n$; в противном случае берут следующую совокупность $n+1$ случайных чисел на интервале (0, 1).

Очевидно, что с ростом числа случайных величин, входящих в систему, процедура их моделирования становится все более громоздкой. Поэтому обычно для системы случайных величин $X_1, X_2, ..., X_n$ используют *приближенный метод моделирования*. Он состоит в моделировании вместо $X_1, X_2, ..., X_n$ таких случайных величин $Z_1, Z_2, ..., Z_n$, которые обладают следующими свойствами:

$$M(Z_i) = M(X_i), \; K(Z_i, Z_j) = K(X_i, X_j),$$

где $M(Z_i)$, $K(Z_i, Z_j)$ — соответственно математическое ожидание Z_i и корреляционный момент величин Z_i и Z_j, $(i, j=1, 2, ..., n)$, т. е. методом приближенного моделирования исходная система случайных величин $X_1, X_2, ..., X_n$ воспроизводится только в рамках «математических ожиданий и корреляционных моментов (или парных коэффициентов корреляции)». Очевидно, что этим методом случайные величины $X_1, X_2, ..., X_n$ воспроизводятся точно в «рамках их закона распределения» только в том случае, если закон распределения величин $X_1, X_2, ..., X_n$ полностью определяется их математическими ожиданиями и парными коэффициентами корреляции (таковым является нормальный закон).

Итак, допустим, что нужно получать значения случайных величин $X_1, X_2, ..., X_n$, если известны математические ожидания $M(X_i) = a_i (i=1, 2, ..., n)$ и корреляционные моменты $K(X_i, X_j) = K_{ij} (ij=1, 2, ..., n)$. Напомним, что $K(X_i, X_j) = M[(X_i - a_i)(X_j - a_j)]$. Пусть $R_1, R_2, ..., R_n$ — независимые случайные величины, равномерно распределенные

на интервале (0, 1). Рассмотрим случайные величины Z_1, Z_2,..., Z_n:

$$
\begin{aligned}
Z_1 &= c_{11}(R_1 - 0{,}5) + a_1,\\
Z_2 &= c_{21}(R_1 - 0{,}5) + c_{22}(R_2 - 0{,}5) + a_2, \qquad (13.2)\\
Z_3 &= c_{31}(R_1 - 0{,}5) + c_{32}(R_2 - 0{,}5) + c_{33}(R_3 - 0{,}5) + a_3,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
Z_n &= c_{n1}(R_1 - 0{,}5) + c_{n2}(R_2 - 0{,}5) + \cdots + c_{nn}(R_n - 0{,}5) + a_n,
\end{aligned}
$$

где c_{ij} $(i, j = 1,..., n)$ — некоторые постоянные коэффициенты.

Опираясь на свойства математического ожидания и учитывая, что $M(R_i) = 1/2$, получаем $M(Z_i) = a_i$ $(i = 1, 2,..., n)$. Действительно, например, $M(Z_1) = M[c_{11}(R_1 - 0{,}5) + a_1] = M[c_{11}(R_1 - 0{,}5)] + Ma_1 = c_{11}[M(R_1) - 0{,}5] + a_1 = a_1$.

Вычислим корреляционные моменты $K(Z_i, Z_j)$ $(i, j = 1,..., n)$, учитывая при этом, что $\sigma^2_{R_i} = M(R_i - 0{,}5)^2 = 1/12$ $(i = 1, 2,..., n)$; а $K(R_i, R_j) = M[(R_i - 0{,}5)(R_j - 0{,}5)] = 0$, так как случайные на интервале (0, 1) числа независимы:

$$
\begin{aligned}
K(Z_1, Z_1) &= M[(Z_1 - M(Z_1))(Z_1 - M(Z_1))] =\\
&= M[c_{11}(R_1 - 0{,}5)\, c_{11}(R_1 - 0{,}5)] = c_{11}^2 M(R_1 - 0{,}5)^2 =\\
&= c_{11}^2 \sigma^2_{R_1} = c_{11}^2/12,\\
K(Z_1, Z_2) &= M[(Z_1 - M(Z_1))(Z_2 - M(Z_2))] =\\
&= M\{c_{11}(R_1 - 0{,}5)[c_{21}(R_1 - 0{,}5) + c_{22}(R_2 - 0{,}5)]\} =\\
&= M[c_{11} c_{21}(R_1 - 0{,}5)^2] + M[c_{11} c_{22}(R_1 - 0{,}5)(R_2 - 0{,}5)] =\\
&= c_{11} c_{21} M(R_1 - 0{,}5)^2 + c_{11} c_{22} M[(R_1 - 0{,}5)(R_2 - 0{,}5)] =\\
&= c_{11} c_{21} \sigma^2_{R_1} + c_{11} c_{22} K(R_1 R_2) = c_{11} c_{21}/12,\\
K(Z_1, Z_2) &= M\{c_{11}(R_1 - 0{,}5)[c_{31}(R_1 - 0{,}5) +\\
&+ c_{32}(R_2 - 0{,}5) + c_{33}(R_3 - 0{,}5)]\} = c_{11} c_{31}/12,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
K(Z_1, Z_n) &= c_{11} c_{n1}/12,\\
K(Z_2, Z_2) &= (c_{21}^2 + c_{22}^2)/12,\\
K(Z_2, Z_3) &= (c_{21} c_{31} + c_{22} c_{32})/12,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
K(Z_n, Z_n) &= (c_{n1}^2 + c_{n2}^2 + \cdots + c_{nn}^2)/12.
\end{aligned}
$$

Приравняв корреляционные моменты $K(Z_i, Z_j)$ случайных величин Z_i и Z_j $(i, j=1, 2,..., n)$ известным значениям корреляционных моментов величин X_i и X_j $(i, j=1, 2,..., n)$, получим уравнения для определения значений c_{ij} $(i, j=1, 2,..., n)$:

$$K_{11} = c_{11}^2/12 \to c_{11} = \sqrt{12K_{11}},$$

$$K_{12} = c_{11}c_{21}/12 \to c_{21} = 12K_{12}/c_{11} = 12K_{12}/\sqrt{12K_{11}},$$

$$K_{13} = c_{11}c_{31}/12 \to c_{31} = 12K_{13}/\sqrt{12K_{11}},$$

. . . . . .

$$K_{1n} = c_{11}c_{n1}/12 \to c_{n1} = \frac{12K_{1n}}{\sqrt{12K_{11}}},$$

$$K_{22} = (c_{21}^2 + c_{22}^2)/12 \to c_{22} = \sqrt{12K_{22} - c_{21}^2} =$$

$$= \sqrt{12K_{22} - \frac{12K_{12}^2}{K_{11}}},$$

$$K_{23} = (c_{21}c_{31} + c_{22}c_{32})/12 \to c_{32} = \frac{12K_{23} - 12K_{12}K_{13}/K_{11}}{\sqrt{12K_{22} - 12K_{12}^2/K_{11}}}$$ и т. д.

После вычисления c_{ij} $(i, j=1, 2,..., n)$ приступают к определению значений случайных величин $X_1, X_2,..., X_n$. Для этого берут n случайных на интервале (0, 1) чисел и, подставляя их в формулы (13.2), получают значения случайных величин.

Коротко остановимся на вопросе формирования реализаций случайных функций. В § 12.1 было показано, что случайную функцию можно рассматривать как совокупность случайных величин — ординат функции при различных допустимых значениях ее аргумента. Поэтому формирование реализации случайной функции $X(t)$ сводится к моделированию системы случайных величин $[X(t_1), X(t_2),..., X(t_n)]$ для некоторой последовательности $t_1, t_2,..., t_n$ значений аргумента t и задача моделирования случайной функции в этом случае ничем не отличается от задачи моделирования системы n случайных величин.

Рассмотрим моделирование случайных функций, наиболее часто встречающихся в практических приложениях, а именно цепей Маркова и стационарных случайных функций. Допустим, что требуется построить реализацию простой однородной цепи Маркова $X(t)$ (см. § 12.9). Предположим, что $X(t)$ принимает конечное множество значений $x_0, x_1, \ldots, x_n$; аргумент t принимает конечное множество равноудаленных друг от друга значений $t_0, t_1, \ldots, t_m$; матрица вероятностей перехода из состояния x_q $(q=1, 2, \ldots, n)$ в состояние $x_{q'}$ $(q'=1, 2, \ldots, n)$ за промежуток времени $\tau=t_i-t_{i-1}$ $(i=1, 2, \ldots, m)$ имеет вид

$$\Gamma(\tau)=\begin{vmatrix} p_{00}, & p_{01}, & \ldots, & p_{0n} \\ p_{10}, & p_{11}, & \ldots, & p_{1n} \\ \cdot & \cdot & \cdot & \cdot \\ p_{n0}, & p_{n1}, & \ldots, & p_{nn} \end{vmatrix};$$

вектор вероятностей состояний в начальный момент

$$\mathrm{p}(t_0)=[p_0(t_0), p_1(t_0), \ldots, p_n(t_0)].$$

При этих условиях построение реализаций $X(t)$ сводится к моделированию системы дискретных случайных величин $X(t_0), X(t_1), \ldots, X(t_m)$, каждая из которых принимает значение из множества $\{x_0, x_1, \ldots, x_n\}$; при этом закон распределения $X(t_0)$ имеет вид

$X(t_0)$	x_0	x_1	$\ldots$	x_n
P	$p_0(t_0)$	$p_1(t_0)$	$\ldots$	$p_n(t_0)$

а условный закон распределения случайной величины $X(t_i)$ $(i=1, 2, \ldots, m)$ (напомним, что для простой цепи Маркова этот закон зависит только от ее состояния в момент t_{i-1}) при условии, что величина $X(t_{i-1})$ приняла значение x_q $(q=0, 1, \ldots, n)$, определяется q-й строкой матрицы $\Gamma(\tau)$, т. е. имеет вид

$X(t_i)$	x_0	x_1	$\ldots$	x_n
P	p_{q0}	p_{q1}	$\ldots$	p_{qn}

Реализации строят так:

1) берут случайное число R на интервале (0, 1) и, используя процедуру моделирования дискретной случай-

ной величины, определяют, какое значение принимает величина $X(t_0)$, т. е. в каком состоянии цепь Маркова находится в начальный момент (допустим, что $X(t_0)=x_i$);

2) берут следующее случайное число R_1 на интервале (0, 1) и, зная условный закон распределения случайной величины $X(t_1)$ при условии, что $X(t_0)=x_i$,

$X(t_1)$	x_0	x_1	...	x_n
P	p_{i0}	p_{i1}	...	p_{in}

определяют, какое значение принимает случайная величина $X(t_1)$; допустим, что $X(t_1)=x_{i_1}$;

3) выбирают случайное число R_2 и, зная условный закон распределения

$X(t_2)$	x_0	x_1	...	x_n
P	p_{i1}	p_{i1}	...	p_{i1^n}

определяют, какое значение примет случайная величина $X(t_2)$; допустим, $X(t_2)=x_{i_2}$, и т. д. вплоть до определения значения случайной величины $X(t_m)$ (допустим, что $X(t_m)=x_{i_m}$).

Таким образом, в результате построения получена следующая реализация простой однородной цепи Маркова: $x_i \to x_{i_1} \to x_{i_2} \to \ldots \to x_{i_m}$. При повторении описанной процедуры с новыми случайными числами формируют другие реализации цепи Маркова.

Если изучают стационарную случайную функцию $X(t)$, то, как было показано в § 12.4, для ее описания достаточно задать математическое ожидание $m_x(t)=m_x$ и корреляционную функцию $k_x(t, t')=k_x(\Delta t=t'-t)$. Поэтому, если известны математическое ожидание m_x и корреляционная функция $k_x(\Delta t)$, формирование реализаций функции $X(t)$ сводится к моделированию системы случайных величин $[X(t_1), X(t_2),..., X(t_n)]$, где $t_1, t_2,..., t_n$ — некоторая последовательность значений аргумента t; математические ожидания этих величин $M[X(t_i)]=m_x$ $(i=1, 2,..., n)$, а корреляционные моменты $K[X(t_i), X(t_j)]=k_x(t_i, t_j)=k_x(\Delta t=t_j-t_i)$. В этом случае можно использовать приближенный метод моделирования.

Таблица 1

Плотность вероятности нормального распределения

$$f(t)=\frac{1}{\sqrt{2\pi}}e^{-t^2/2}$$

t	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3188	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0,0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Значение функции $\Phi(t)=\frac{2}{\sqrt{2\pi}}\int_0^t e^{-\frac{t^2}{2}}dt$

t	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0080	0,0160	0,0239	0,0319	0,0399	0,0478	0,0558	0,0638	0,0717
0,1	0797	0876	0955	1034	1113	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2960	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4907	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6629	6679	6729	6778
1,0	0,6827	0,6875	0,6923	0,6970	0,7017	0,7063	0,7109	0,7154	0,7199	0,7243
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7994	8029
1,3	8064	8098	8132	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8789	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9090
1,7	9109	9127	9146	9164	9181	9199	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9426	9439	9451	9464	9476	9488	9500	9512	9523	9534
2,0	0,9545	0,9556	0,9566	0,9576	0,9586	0,9596	0,9606	0,9616	0,9625	0,9634
2,1	9643	9651	9660	9668	9676	9684	9692	9700	9707	9715
2,2	9722	9729	9736	9743	9749	9756	9762	9768	9774	9780

Продолжение табл. 2

t	0	1	2	3	4	5	6	7	8	9
2,3	9786	9791	9797	9802	9807	9812	9817	9822	9827	9832
2,4	9836	9841	9845	9849	9853	9857	9861	9865	9869	9872
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
2,6	9907	9910	9912	9915	9917	9920	9922	9924	9926	9928
2,7	9931	9933	9935	9937	9939	9940	9942	9944	9946	9947
2,8	9949	9951	9952	9953	9955	9956	9958	9959	9960	9961
2,9	9963	9964	9965	9966	9967	9968	9969	9970	9971	9972
3,0	0,9973	0,9974	0,9975	0,9976	0,9976	0,9977	0,9978	0,9979	0,9979	0,9980
3,1	9981	9981	9982	9983	9983	9984	9984	9985	9985	9986
3,2	9986	9987	9987	9988	9988	9989	9989	9989	9990	9990
3,3	9990	9991	9991	9991	9992	9992	9992	9992	9993	9993
3,4	9993	9994	9994	9994	9994	9994	9995	9995	9995	9995
3,5	9995	9996	9996	9996	9996	9996	9996	9996	9997	9997
3,6	9997	9997	9997	9997	9997	9997	9997	9998	9998	9998
3,7	9998	9998	9998	9998	9998	9998	9998	9998	9998	9998
3,8	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999
3,9	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999
4,0	0,999936	9999	9999	9999	9999	9999	9999	9999	9999	9999
4,5	0,999994	—	—	—	—	—	—	—	—	—
5,0	0,99999994	—	—	—	—	—	—	—	—	—

Таблица 3

Значения случайных чисел равномерно распределенных на интервале (0, 1)*

10 097	32 533	76 520	13 586	34 673	54 876
37 542	04 805	64 894	74 296	24 805	24 037
08 422	68 953	19 645	09 303	23 209	02 560
99 019	02 529	09 376	70 715	38 311	31 165
12 807	99 970	80 157	36 147	64 032	36 653
80 959	09 117	39 292	74 945	66 065	74 717
20 636	10 402	00 822	91 665	31 060	10 805
15 953	34 764	35 080	33 606	85 269	77 602
88 676	74 397	04 436	27 659	63 573	32 135
98 951	16 877	19 171	76 833	73 796	45 753
34 072	76 850	36 697	36 170	65 813	39 885
45 571	82 406	35 303	42 614	86 779	07 439
02 051	65 692	68 665	74 818	73 053	85 247
05 325	47 048	90 553	57 548	28 468	28 709
03 529	64 778	35 808	34 282	60 935	20 344
11 199	29 170	98 520	17 767	14 905	68 607
23 403	09 732	11 805	05 431	39 808	27 732
18 623	88 579	83 452	99 634	06 288	98 083
83 491	25 624	88 685	40 200	86 507	58 401
35 273	88 435	99 594	67 348	87 517	64 969
22 109	40 555	60 970	93 433	50 500	73 998
50 725	68 248	29 405	24 201	52 775	67 851
13 746	70 078	18 475	40 610	68 711	77 817
36 766	67 951	90 364	76 493	29 609	11 062
91 826	08 928	93 785	61 368	23 478	34 113
65 481	17 674	17 468	50 950	79 335	51 748
80 124	35 635	17 727	08 015	82 391	90 324
74 350	99 817	77 402	77 214	50 024	23 356
69 915	26 803	66 252	29 148	24 892	09 994
09 893	20 505	14 225	68 514	83 647	76 938

* Все значения, приведенные в таблице, увеличены в 10^5 раз.

Таблица 4

F-распределение для уровня значимости $\alpha = 0{,}05$

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	11
1	161	200	216	225	230	234	237	239	241	242	243
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	2,31
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	2,22
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,27	2,21	2,16	2,12
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,07	2,04
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,02	1,98
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,01	1,97	1,93
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97	1,92	1,88
200	3,89	3,04	2,65	2,41	2,26	2,14	2,05	1,98	1,92	1,87	1,83
400	3,86	3,02	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,81
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83	1,79

Продолжение табл. 4

k_2 \ k_1	12	13	16	20	30	40	50	75	100	200	500	∞
1	244	245	246	248	250	251	252	253	253	254	254	254
2	19,41	19,42	19,43	19,44	19,46	19,47	19,47	19,48	19,49	19,49	19,50	19,50
3	8,74	8,71	8,69	8,66	8,62	8,60	8,58	8,57	8,56	8,54	8,54	8,53
4	5,91	5,87	5,84	5,80	5,74	5,71	5,70	5,68	5,66	5,65	5,64	5,63
5	4,68	4,64	4,60	4,56	4,50	4,46	4,44	4,42	4,40	4,38	4,38	4,36
6	4,00	3,96	3,92	3,87	3,81	3,77	3,75	3,72	3,71	3,69	3,68	3,67
7	3,57	3,52	3,49	3,44	3,38	3,34	3,32	3,29	3,28	3,25	3,24	3,23
8	3,28	3,23	3,20	3,15	3,08	3,05	3,03	3,00	2,98	2,96	2,94	2,93
9	3,07	3,02	2,98	2,93	2,86	2,82	2,80	2,77	2,76	2,73	2,72	2,71
10	2,91	2,86	2,82	2,77	2,70	2,67	2,64	2,61	2,59	2,56	2,55	2,54
11	2,79	2,74	2,70	2,65	2,57	2,53	2,50	2,47	2,45	2,42	2,41	2,40
12	2,69	2,64	2,60	2,54	2,46	2,42	2,40	2,36	2,35	2,32	2,31	2,30
14	2,53	2,48	2,44	2,39	2,31	2,27	2,24	2,21	2,19	2,16	2,14	2,13
16	2,42	2,37	2,33	2,28	2,20	2,16	2,13	2,09	2,07	2,04	2,02	2,01
20	2,28	2,23	2,18	2,12	2,04	1,99	1,96	1,92	1,90	1,87	1,85	1,84
24	2,18	2,13	2,09	2,02	1,94	1,89	1,86	1,82	1,80	1,76	1,74	1,73
30	2,09	2,04	1,99	1,93	1,84	1,79	1,76	1,72	1,69	1,66	1,64	1,62
40	2,00	1,95	1,90	1,84	1,74	1,69	1,66	1,61	1,59	1,55	1,53	1,51
50	1,95	1,90	1,85	1,78	1,69	1,63	1,60	1,55	1,52	1,48	1,46	1,44
70	1,89	1,84	1,79	1,72	1,62	1,56	1,53	1,47	1,45	1,40	1,37	1,35
100	1,85	1,79	1,75	1,68	1,57	1,51	1,48	1,42	1,39	1,34	1,30	1,28
200	1,80	1,74	1,69	1,62	1,52	1,45	1,42	1,35	1,32	1,26	1,22	1,19
400	1,78	1,72	1,67	1,60	1,49	1,42	1,38	1,32	1,28	1,22	1,16	1,13
∞	1,75	1,69	1,64	1,57	1,46	1,40	1,35	1,28	1,24	1,17	1,11	1,00

Таблица 5

Значения t-распределения Стьюдента $P = \int_{-t}^{t} S(t, k)dt$

k \ P	0,95	0,99	0,999	k \ P	0,95	0,99	0,999
4	2,78	4,60	8,61	19	2,093	2,861	3,883
5	2,57	4,03	6,86	20	2,086	2,845	3,849
6	2,45	3,71	5,96	25	2,064	2,797	3,745
7	2,37	3,50	5,41	30	2,045	2,756	3,659
8	2,31	3,36	5,04	35	2,032	2,729	3,600
9	2,26	3,25	4,78	40	2,023	2,708	3,558
10	2,23	3,17	4,59	45	2,016	2,692	3,527
11	2,20	3,11	4,44	50	2,009	2,679	3,502
12	2,18	3,06	4,32	60	2,001	2,662	3,464
13	2,16	3,01	4,22	70	1,996	2,649	3,439
14	2,15	2,98	4,14	80	1,991	2,640	3,418
15	2,13	2,95	4,07	90	1,987	2,633	3,403
16	2,12	2,92	4,02	100	1,984	2,627	3,392
17	2,11	2,90	3,97	120	1,980	2,617	3,374
18	2,10	2,88	3,92	∞	1,960	2,576	3,291

Таблица 6

Значения t-распределения Стьюдента $P = \int_{-\infty}^{t} S(t, k)dt$

k \ P	0,95	0,99	0,999	k \ P	0,95	0,99	0,999
5	3,04	5,04	9,43	20	2,145	2,932	3,979
6	2,78	4,36	7,41	25	2,105	2,852	3,819
7	2,62	3,96	6,37	30	2,079	2,802	3,719
8	2,51	3,71	5,73	35	2,061	2,768	3,652
9	2,43	3,54	5,31	40	2,048	2,742	3,602
10	2,37	3,41	5,01	45	2,038	2,722	3,565
11	2,33	3,31	4,79	50	2,030	2,707	3,532
12	2,29	3,23	4,62	60	2,018	2,683	3,492
13	2,26	3,17	4,48	70	2,009	2,667	3,462
14	2,24	3,12	4,37	80	2,003	2,655	3,439
15	2,22	3,08	4,28	90	1,998	2,646	3,423
16	2,20	3,04	4,20	100	1,994	2,639	3,409
17	2,18	3,01	4,13	∞	1,960	2,576	3,291
18	2,17	2,98	4,07				

Таблица 7

Значения χ^2 в зависимости от вероятности $P\left(\alpha^2 > \chi^2_\alpha\right)$ и числа степеней свободы k

Число степеней свободы k	Вероятность															
	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,005	0,002	0,001
1	0,00016	0,0006	0,0039	0,016	0,064	0,148	0,455	1,07	1,64	2,7	3,8	5,4	6,6	7,9	9,5	10,83
2	0,020	0,040	0,103	0,211	0,446	0,713	1,386	2,41	3,22	4,6	6,0	7,8	9,2	11,6	12,4	13,8
3	0,115	0,185	0,352	0,584	1,005	1,424	2,366	3,66	4,64	6,3	7,8	9,8	11,3	12,8	14,8	16,3
4	0,30	0,43	0,71	1,06	1,65	2,19	3,36	4,9	6,0	7,8	9,5	11,7	13,3	14,9	16,9	18,5
5	0,55	0,75	1,14	1,61	2,34	3,00	4,35	6,1	7,3	9,2	11,1	13,4	15,1	16,3	18,9	20,5
6	0,187	1,13	1,63	2,20	3,07	3,83	5,35	7,2	8,6	10,6	12,6	15,0	16,8	18,6	20,7	22,5
7	1,24	1,56	2,17	2,83	3,82	4,67	6,34	8,4	9,8	12,0	14,1	16,6	18,5	20,3	22,6	24,4
8	1,65	2,03	2,73	3,49	4,59	5,53	7,34	9,5	11,0	13,4	15,5	18,2	20,1	21,9	24,3	21,6
9	2,09	2,53	3,32	4,17	5,38	6,39	8,35	10,7	12,2	14,7	16,9	19,7	21,7	23,6	26,1	27,9
10	2,56	3,06	3,94	4,86	6,18	7,27	9,34	11,8	13,4	16,0	18,3	21,2	23,2	25,2	27,7	29,6
11	3,1	3,6	4,6	5,6	7,0	8,1	10,3	12,9	14,6	17,3	19,7	22,6	24,7	26,8	29,4	31,3
12	3,6	4,2	5,2	6,3	7,8	9,0	11,3	14,0	15,8	18,5	21,0	24,1	26,2	28,3	31,0	32,9
13	4,1	4,8	5,9	7,0	8,6	9,9	12 3	15,1	17,0	19,8	22,4	25,5	27,7	29,8	32,5	34,5
14	4,7	5,4	6,6	7,8	9,5	10,8	13,3	16,2	18,2	21,1	23,7	26,9	29,1	31,0	34,0	36,1
15	5,2	6,0	7,3	8,5	10,3	11,7	14,3	17,3	19,3	22,3	25,0	28,3	30,6	32,5	35,5	37,7
16	5,8	6,6	8,0	9,3	11,2	12,6	15,3	18,4	20,5	23,5	26,3	29,6	32,0	34,0	37,0	39,2
17	6,4	7,3	8,7	10,1	12,0	13,5	16,3	19,5	21,6	24,8	27,6	31,0	33,4	35,5	38,5	40,8
18	7,0	7,9	9,4	10,9	12,9	14,4	17,3	20,6	22,8	26,0	28,9	32,3	34,8	37,0	40,0	42,3

Продолжение табл. 7

Число степеней свободы k	Вероятность															
	0,99	0,98	0,95	0,90	0,80	0,70	0,60	0,30	0,20	0,10	0,05	0,02	0,01	0,005	0,002	0,001
19	7,6	8,6	10,1	11,7	13,7	15,4	18,3	21,7	23,9	27,2	30,1	33,7	36,2	38,5	41,5	43,8
20	8,3	9,2	10,9	12,4	14,6	16,3	19,3	22,8	25,0	28,4	31,4	35,0	37,6	40,0	43,0	45,3
21	8,9	9,9	11,6	13,2	15,4	17,2	20,3	23,9	26,2	29,6	32,7	36,3	38,9	41,5	44,5	46,8
22	9,5	10,6	12,3	14,0	16,3	18,1	21,3	24,9	27,3	30,8	33,9	37,7	40,3	42,5	46,0	48,3
23	10,2	11,3	13,1	14,8	17,2	19,0	22,3	26,0	28,4	32,0	35,2	39,0	41,6	44,0	47,5	49,7
24	10,9	12,0	13,8	15,7	18,1	19,9	23,3	27,1	29,6	33,2	36,4	40,3	43,0	45,5	48,5	51,2
25	11,5	12,7	14,6	16,5	18,9	20,9	24,3	28,1	30,7	34,4	37,7	41,6	44,3	47,0	50,0	52,6
26	12,2	13,4	15,4	17,3	19,8	21,8	25,3	29,3	31,8	35,6	38,9	42,9	45,6	48,0	51,5	54,1
27	12,9	14,1	16,2	18,1	20,7	22,7	26,3	30,3	32,9	36,7	40,1	44,1	47,0	49,5	53,0	55,5
28	13,6	14,8	16,9	18,9	21,6	23,6	27,3	31,4	34,0	37,9	41,3	45,4	48,3	51,0	54,5	56,9
29	14,3	15,6	17,7	19,8	22,5	24,6	28,3	32,5	35,1	39,1	42,6	46,7	49,6	52,5	56,0	58,3
30	15,0	16,3	18,5	20,6	23,4	25,5	29,3	33,5	36,3	40,3	43,8	48,0	50,9	54,0	57,5	59,7