N.Sh. Kremer

PROBABILITY THEORY AND MATHEMATICAL STATISTICS

Second Edition

Textbook

ЮНИТИ
UNITY
Moscow • 2004

Н.Ш. Кремер

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Второе издание,
переработанное и дополненное

Рекомендовано Министерством образования Российской Федерации в качестве ***учебника*** *для студентов высших учебных заведений, обучающихся по экономическим специальностям*

ЮНИТИ
UNITY
Москва • 2004

УДК 591.2(075.8)
ББК 22.17я73
К79

Рецензенты:
кафедра математической статистики и эконометрики Московского государственного университета экономики, статистики и информатики (МЭСИ)
(зав. кафедрой д-р экон. наук, проф. В.С. Мхитарян);
д-р физ.-мат. наук, проф. *В.Ф. Гапошкин*;
канд. техн. наук, доц. *Г.Л. Эпштейн*

Главный редактор издательства
доктор экономических наук *Н.Д. Эриашвили*

Кремер Н.Ш.

К79 Теория вероятностей и математическая статистика: Учебник для вузов. — 2-е изд., перераб. и доп.— М.: ЮНИТИ-ДАНА, 2004. — 573 с.

ISBN 5-238-00573-3

Это не только учебник, но и краткое руководство к решению задач. Излагаемые основы теории вероятностей и математической статистики сопровождаются большим количеством задач (в том числе экономических), приводимых с решениями и для самостоятельной работы. При этом упор делается на основные понятия курса, их теоретико-вероятностный смысл и применение. Приводятся примеры использования вероятностных и математико-статистических методов в задачах массового обслуживания и моделях финансового рынка.

Для студентов и аспирантов экономических специальностей и направлений, а также преподавателей вузов, научных сотрудников и экономистов.

ББК 22.17я73

ISBN 5-238-00573-3

Оглавление

Предисловие

Издательство ЮНИТИ-ДАНА продолжает выпуск учебников и учебных пособий по математическим дисциплинам для студентов и абитуриентов экономических вузов.

Мотивацией подготовки данного учебника явилось также то, что в настоящее время ощущается нехватка доступных для студентов-экономистов учебников по дисциплине «теория вероятностей и математическая статистика». В первую очередь это касается студентов, обучающихся в вузе без отрыва от производства, для многих из которых учебник служит основным источником учебной информации. Вышедшие из печати в последнее время учебники и пособия по теории вероятностей и математической статистике ориентированы в основном на студентов технических вузов и предполагают достаточно высокий уровень их математической подготовки.

Данный учебник написан в соответствии с требованиями Государственного образовательного стандарта второго поколения и Примерной программой дисциплины «Математика», утвержденной Минобразованием РФ в 2000 г. Основной принцип, которым руководствовался автор при подготовке курса теории вероятностей и математической статистики для экономистов, — **повышение уровня фундаментальной математической подготовки студентов с усилением ее прикладной экономической направленности.**

Учебник состоит из двух разделов, отражающих основы дисциплины: I «Теория вероятностей» (гл. 1 «Основные понятия и теоремы теории вероятностей»); гл. 2 «Повторные независимые испытания»; гл. 3 «Случайные величины»; гл. 4 «Основные законы распределения»; гл. 5 «Многомерные случайные величины»; гл. 6 «Закон больших чисел и предельные теоремы») и II «Математическая статистика» (гл. 8 «Вариационные ряды и их характеристики»; гл. 9 «Основы математической теории выборочного метода»; гл. 10 «Проверка статистических гипотез»; гл. 11 «Дисперсионный анализ»; гл. 12 «Корреляционный анализ»; гл. 13 «Регрессионный анализ»; гл. 14 «Введение в анализ временных рядов»). Наряду с этим в учебнике в сжатой форме рассматривается применение вероятностных и математико-статистических методов в решении ряда прикладных экономических задач: в разд. I — это гл. 7 «Элементы теории случайных процессов и теории массового обслуживания» и в разд. II — гл. 15 «Линейные регрессионные модели финансового рынка» (гл. 15 (с. 519—532) написана доц. *Путко Б.А.*).

Известно, что новый учебный материал усваивается студентами (особенно обучающимися без отрыва от производства) значительно легче, если он сопровождается достаточно большим числом иллюстрирующих его примеров. Поэтому автором сделана попытка соединить в одной книге учебник и краткое руководство к решению задач. При подготовке задач были использованы различные пособия и методические материалы. Часть задач составлена автором специально для учебника.

Задачи с решениями (в том числе с экономическим содержанием) рассматриваются на протяжении всего изложения учебного материала. Более сложные, комплексные, а также дополнительные задачи с решениями приводятся в ряде глав в специальном параграфе «Решение задач». Задачи для самостоятельной работы рассматриваются в конце каждой главы в рубрике «Упражнения» (нумерация задач единая — начинается в основном тексте главы и продолжается в этой рубрике). Ответы к этим задачам приводятся в конце книги. Необходимые для решения задач математико-статистические таблицы даются в приложении. В конце книги приводится развернутый предметный указатель основных понятий курса.

Во второе издание включен новый § 2.5 «Полиномиальная схема», являющийся обобщением схемы Бернулли, а также § 7.8 «Понятие о методе статистических испытаний (методе Монте-Карло)» — одном из эффективных методов статистического моделирования. Дополнено изложение некоторых вопросов, например, приводятся свойства условного математического ожидания, дается механическая интерпретация дисперсии случайной величины, рассматривается построение наиболее мощного статистического критерия с помощью леммы Неймана—Пирсона, приводится критерий χ^2 однородности выборок, дается оценка существенности различия характеристик тесноты связи, строится доверительный интервал для дисперсии возмущений в регрессионной модели и т.п. Внесены коррективы в изложение глав 12—14 раздела «Математическая статистика», в частности, с целью приближения его к изучению дисциплины «Эконометрика», впервые включенной в 2000 г. в Государственный образовательный стандарт большинства экономических специальностей. Добавлены новые задачи с решениями и для самостоятельной работы. Исправлены замеченные опечатки и неточности.

Автор выражает глубокую благодарность проф. *В.С. Мхитаряну*, проф. *В.Ф. Гапошкину* и доц. *Г.Л. Эпштейну* за рецензирование рукописи и сделанные ими замечания.

В книге знаком □ обозначается начало доказательства теоремы, знаком ■ — ее окончание; знаком ▷ — начало условия задачи, знаком ▶ — окончание ее решения.

Введение

Задача любой науки, в том числе экономической, состоит в выявлении и исследовании закономерностей, которым подчиняются реальные процессы. Найденные закономерности, относящиеся к экономике, имеют не только теоретическую ценность, они широко применяются на практике — в планировании, управлении и прогнозировании.

Теория вероятностей *— математическая наука, изучающая закономерности случайных явлений.* Под случайными явлениями понимаются явления с неопределенным исходом, происходящие при неоднократном воспроизведении определенного комплекса условий.

Очевидно, что в природе, технике и экономике нет явлений, в которых не присутствовали бы элементы случайности. Существуют два подхода к изучению этих явлений. Один из них — классический, или «детерминистский», состоит в том, что выделяются основные факторы, определяющие данное явление, а влиянием множества остальных, второстепенных, факторов, приводящих к случайным отклонениям его результата, пренебрегают. Таким образом выявляется основная закономерность, свойственная данному явлению, позволяющая однозначно предсказать результат по заданным условиям. Этот подход часто используется в естественных («точных») науках.

При исследовании многих явлений и прежде всего социально-экономических такой подход неприемлем. В этих явлениях необходимо учитывать не только основные факторы, но и множество второстепенных, приводящих к случайным возмущениям и искажениям результата, т.е. вносящих в него элемент неопределенности. Поэтому другой подход к изучению явлений состоит в том, что элемент неопределенности, свойственный случайным явлениям и обусловленный второстепенными факторами, требует специальных методов их изучения. Разработкой таких методов, изучением специфических закономерностей, наблюдаемых в случайных явлениях, и занимается теория вероятностей.

Математическая статистика *— раздел математики, изучающий математические методы сбора, систематизации, обработки и интерпретации результатов наблюдений с целью выявления статистических закономерностей.* Математическая статистика опирается на теорию вероятностей. Если теория вероятностей изучает закономерности случайных явлений на основе абстрактного описания действительности (теоретической вероятностной модели), то математическая статистика оперирует непосредственно результатами наблюдений над случайным явлением, представляющими выборку из некоторой конечной или гипотетической бесконечной генеральной совокупности. Используя результаты, полученные теорией вероятностей, математическая статистика позволяет не только оценить значения искомых характеристик, но и выявить степень точности получаемых при обработке данных выводов.

Если говорить кратко, теория вероятностей позволяет находить вероятности «сложных» событий через вероятности «простых» событий (связанных с ними каким-либо образом), а математическая статистика по наблюденным значениям (выборке) оценивает вероятности этих событий либо осуществляет проверку предположений (гипотез) относительно этих вероятностей.

Изучение вероятностных моделей дает возможность понять различные свойства случайных явлений на абстрактном и обобщенном уровне, не прибегая к эксперименту. В математической статистике, наоборот, исследование связано с конкретными данными и идет от практики (наблюдения) к гипотезе и ее проверке.

При большом числе наблюдений случайные воздействия в значительной мере погашаются (нейтрализуются) и получаемый результат оказывается практически неслучайным, предсказуемым. Это утверждение (принцип) и является базой для практического использования вероятностных и математико-статистических методов исследования. Цель указанных методов состоит в том, чтобы, минуя сложное (а зачастую и невозможное) исследование отдельного случайного явления, изучить закономерности массовых случайных явлений, прогнозировать их характеристики, влиять на ход этих явлений, контролировать их, ограничивать область действия случайности.

Первые работы, в которых зарождались основные понятия теории вероятностей, появились в XVI—XVII вв. Они принадлежали Д. Кардано, Б. Паскалю, П. Ферма, Х. Гюйгенсу и др. и представляли попытки создания теории азартных игр с целью

дать рекомендации игрокам. Следующий этап развития теории вероятностей связан с именем Я. Бернулли (XVII — начало XVIII в.), который доказал теорему, теоретически обосновавшую накопленные ранее факты и названную в дальнейшем «законом больших чисел».

Дальнейшее развитие теории вероятностей приходится на XVII—XIX вв. благодаря работам А. Муавра, П. Лапласа, К. Гаусса, С. Пуассона и др. Весьма плодотворный период развития «математики случайного» связан с именами русских математиков П.Л. Чебышева, А.М. Ляпунова и А.А. Маркова (XIX — начало XX в.).

Большой вклад в последующее развитие теории вероятностей и математической статистики внесли российские математики С.Н. Бернштейн, В.И. Романовский, А.Н. Колмогоров, А.Я. Хинчин, Ю.В. Линник, Б.В. Гнеденко, Н.В. Смирнов, Ю.В. Прохоров и др., а также ученые англо-американской школы Стьюдент (псевдоним В. Госсета), Р. Фишер, Э. Пирсон, Е. Нейман, А. Вальд и др. Особо следует отметить неоценимый вклад академика А.Н. Колмогорова в становление теории вероятностей как математической науки.

Широкому внедрению математико-статистических методов исследования способствовало появление во второй половине XX в. электронных вычислительных машин и, в частности, персональных компьютеров. Статистические программные пакеты сделали эти методы более доступными и наглядными, так как трудоемкую работу по расчету различных статистик, параметров, характеристик, построению таблиц и графиков в основном стал выполнять компьютер, а исследователю осталась главным образом творческая работа: постановка задачи, выбор методов ее решения и интерпретация результатов.

Появление мощных и удобных статистических пакетов для персональных компьютеров позволяет использовать их не только как специальный инструмент научных исследований, но и как общеупотребительный инструмент плановых, аналитических, маркетинговых отделов производственных и торговых корпораций, банков и страховых компаний, правительственных и медицинских учреждений и даже представителей мелкого бизнеса. Среди множества используемых для этих целей пакетов прикладных программ выделим популярные в России универсальные и специализированные статистические пакеты: отечественные STADIA, Эвриста, Статистик-консультант, Олимп: СтатЭксперт и американские STATGRAPHICS, SPSS, SYSTAT, STATISTICA/w и др.

Раздел I
Теория вероятностей

Глава 1 Основные понятия и теоремы теории вероятностей

1.1. Классификация событий

Одним из основных понятий теории вероятностей является понятие события.

Случайным событием (*возможным событием* или просто *событием*) называется любой факт, который в результате испытания может произойти или не произойти.

Под *испытанием* (*опытом*, *экспериментом*) в этом определении понимается выполнение определенного комплекса условий, в которых наблюдается то или иное явление, фиксируется тот или иной результат. Испытание (опыт) может быть осуществлено человеком, но может проводиться и независимо от человека, выступающего в этом случае в роли наблюдателя.

Приведем примеры событий.

1. Появление герба (реверса — оборотной стороны) при подбрасывании монеты.
2. Выигрыш автомобиля по билету денежно-вещевой лотереи.
3. Выход бракованного изделия с конвейера предприятия.
4. Выпадение более 1000 мм осадков в данном географическом пункте за определенный год.

Событие — это не какое-нибудь происшествие, а лишь возможный *исход*, результат испытания (опыта, эксперимента). События обозначаются прописными (заглавными) буквами латинского алфавита: A, B, C.

Если при каждом испытании, при котором происходит событие A, происходит и событие B, то говорят, что *A влечет за собой событие B* (*входит* в B, является *частным случаем*, *вариантом* B) или *B включает событие A*, и обозначают $A \subset B$. Например, если событие A — изделие 1-го сорта, B — изделие 2-го сорта, C — изделие стандартное, то $A \subset C$ и $B \subset C$.

Если одновременно $A \subset B$ и $B \subset A$, то в этом случае события A и B называют *равносильными* и обозначают $A = B$.

События называются *несовместными* (*несовместимыми*), если наступление одного из них исключает наступление любого другого. В противном случае события называются *совместными* (*совместимыми*). Например, выигрыш по одному билету денежно-вещевой лотереи двух ценных предметов — события несовместные, а выигрыш тех же предметов по двум билетам — события совместные. Получение студентом на экзамене по одной дисциплине оценок «отлично», «хорошо» и «удовлетворительно» — события несовместные, а получение тех же оценок на экзаменах по трем дисциплинам — события совместные.

Событие называется *достоверным* (обозначаем буквой Ω), если в результате испытания оно обязательно должно произойти

Событие называется *невозможным* (обозначаем символом $\varnothing$), если в результате испытания оно вообще не может произойти. Например, если в партии все изделия стандартные, то извлечение из нее стандартного изделия — событие достоверное, а извлечение при тех же условиях бракованного изделия — событие невозможное.

События называются *равновозможными*, если в результате испытания по условиям симметрии ни одно из этих событий не является объективно более возможным. Например, извлечение туза, валета, короля или дамы из колоды карт либо появление герба или решки при подбрасывании монеты — события равновозможные. Так, если монета «правильная», выполнена симметрично, то нет никаких оснований считать «появление герба» при подбрасывании монеты событием объективно более возможным, чем «появление решки».

Равновозможные события не могут появляться иначе, чем в испытаниях, обладающих симметрией возможных исходов; и наше незнание того, какое из событий объективно более возможно при отсутствии симметрии исходов, не может служить основанием, чтобы считать события равновозможными.

Несколько событий называются *единственно возможными*, если в результате испытания обязательно должно произойти хотя бы одно из них. Например, события, состоящие в том, что в семье из двух детей: A — «два мальчика», B — «один мальчик, одна девочка», C — «две девочки» — являются единственно возможными.

Другой пример. События, состоящие в том, что при 10 выстрелах число m попаданий в цель: $D - m < 2$, $E - m < 8$, $F - m > 5$ также являются единственно возможными, так как при любом результате стрельбы обязательно произойдет хотя бы одно из этих событий (например, при $m = 9$ — событие F, при $m = 1$ — событие D или E и т.д.).

Несколько событий образуют *полную группу* (*полную систему*), если они являются единственно возможными и несовместными исходами испытания. Это означает, что в *результате испытания обязательно должно произойти одно и только одно из этих событий.* Так, в приведенных двух последних примерах события A, B, C образуют полную группу, так как они единственно возможные и несовместные, а события D, E, F — полную группу не образуют, так как они только единственно возможные, но совместные[1].

Частным случаем событий, образующих полную группу, являются противоположные события. Два несовместных события, из которых одно должно обязательно произойти, называются *противоположными*[2]. Событие, противоположное событию A, будем обозначать $\overline{A}$.

Например, «появление герба» и «появление решки» при подбрасывании монеты, «отсутствие бракованных изделий» и «наличие хотя бы одного бракованного изделия» в партии — события противоположные.

1.2. Классическое определение вероятности

Для практической деятельности важно уметь сравнивать события по степени возможности их наступления. Очевидно, события: «выпадение дождя» и «выпадение снега» в первый день лета в данной местности, «выигрыш по одному билету» и «выигрыш по каждому из n приобретенных билетов» денежно-вещевой лотереи — обладают разной степенью возможности их наступления. Поэтому для сравнения событий нужна определенная мера.

Численная мера степени объективной возможности наступления события называется *вероятностью события.*

Это определение, качественно отражающее понятие вероятности события, не является математическим. Чтобы оно таким стало, необходимо определить его количественно.

Пусть исходы некоторого испытания образуют полную группу событий и равновозможны, т.е. единственно возможны, несовместны и равновозможны. Такие исходы называются *элементарными исходами, случаями* или *шансами*[1]. При этом говорят, что испытание сводится к *схеме случаев* или *«схеме урн»* (ибо любую вероятностную задачу для рассматриваемого испытания можно заменить эквивалентной задачей с урнами и шарами разных цветов).

Случай называется *благоприятствующим* (*благоприятным*) событию A, если появление этого случая влечет за собой появление события A.

Согласно классическому определению ***вероятность***[2] ***события*** *A равна отношению числа случаев, благоприятствующих ему, к общему числу случаев*, т.е.

$$P(A) = \frac{m}{n}, \qquad \textbf{(1.1)}$$

где $P(A)$— вероятность события A;

m — число случаев, благоприятствующих событию A;

n — общее число случаев.

▷ **Пример 1.1.** При бросании игральной кости возможны шесть исходов — выпадение 1, 2, 3, 4, 5, 6 очков. Какова вероятность появления четного числа очков?

Решение. Все $n = 6$ исходов образуют полную группу событий и равновозможны, т.е. единственно возможны, несовместны и равновозможны. Событию A — «появление четного числа очков» благоприятствуют 3 исхода (случая) — 2, 4 и 6 очков. По формуле **(1.1)**

$$P(A) = 3/6 = 1/2. \ \blacktriangleright$$

Классическое определение (точнее, классическая формула) вероятности **(1.1)** долгое время, с XVII вплоть до XIX в., рассматривалось действительно как определение вероятности, так как в то время методы теории вероятностей применялись в основном к азартным играм, которые сводились к схеме случаев, или в задачах, которые искусственно сводились к этой схеме. В настоящее время формальное определение вероятности не дается (это понятие считается первичным и не определяется, а при его пояснении используют понятие относительной частоты события (см. § 1.3)).

[1] В некоторых курсах теории вероятностей в понятие «полной группы событий» не включается требование несовместности событий. При такой трактовке события D, E, F также будут образовывать полную группу.

[2] В литературе такие события называют также *взаимно-дополнительными.*

[1] В теоретико-множественной трактовке (см. § 1.12) такие исходы называют *элементарными событиями.*

[2] Для вероятности события A в литературе используется также обозначение $Pr(A)$ (сокращение слова *probability* (вероятность)).

Поэтому *классическое определение (классическую формулу) вероятности* **(1.1)** *следует рассматривать не как определение, а как метод вычисления вероятностей для испытаний, сводящихся к схеме случаев.*

Отметим свойства вероятности события.

1. *Вероятность любого события заключена между нулем и единицей*, т.е.

$$0 \le P(A) \le 1. \qquad \textbf{(1.2)}$$

2. *Вероятность достоверного события равна единице*, т.е.

$$P(\Omega) = 1.$$

3. *Вероятность невозможного события равна нулю*, т.е.

$$P(\varnothing) = 0.$$

□ Свойства очевидны, так как $P(A) = m/n$, а число m благоприятствующих случаев для любого события удовлетворяет неравенству $0 \le m \le n$, для достоверного события равно n ($m = n$) и для невозможного события равно нулю ($m = 0$). ■

События, вероятности которых очень малы (близки к нулю) или очень велики (близки к единице), называются соответственно *практически невозможными* или *практически достоверными* событиями.

1.3. Статистическое определение вероятности

Выше отмечено, что классическое определение вероятности применимо только для тех событий, которые могут появиться в результате испытаний, обладающих симметрией возможных исходов, т.е. сводящихся к схеме случаев. Однако существует большой класс событий, вероятности которых не могут быть вычислены с помощью классического определения. В первую очередь это события, которые не являются равновозможными исходами испытания. Например, если монета сплющена, то, очевидно, события «появление герба» и «появление решки» при подбрасывании монеты нельзя считать равновозможными, и формула **(1.1)** для расчета вероятности любого из них окажется неприменима.

Но есть и другой подход при оценке вероятности событий, основанный на том, насколько часто будет появляться данное событие в произведенных испытаниях. В этом случае используется статистическое определение вероятности.

Статистической вероятностью *события A называется относительная частота (частость) появления этого события в n произведенных испытаниях, т.е.*

$$\tilde{P}(A) = w(A) = \frac{m}{n}, \qquad \textbf{(1.3)}$$

где $\tilde{P}(A)$ — статистическая вероятность события A;

$w(A)$ — относительная частота (частость) события A;

m — число испытаний, в которых появилось событие A;

n — общее число испытаний.

В отличие от «математической» вероятности $P(A)$, рассматриваемой в классическом определении **(1.1)**, статистическая вероятность $\tilde{P}(A)$ является характеристикой *опытной, экспериментальной.* Если $P(A)$ есть доля случаев, благоприятствующих событию A, которая определяется непосредственно, без каких-либо испытаний, то $\tilde{P}(A)$ есть доля тех фактически произведенных испытаний, в которых событие A появилось.

Статистическое определение вероятности, как и понятия и методы теории вероятностей в целом, применимы не к любым событиям с неопределенным исходом, которые в житейской практике считаются случайными, а только к тем из них, которые обладают определенными свойствами.

1. Рассматриваемые события должны быть *исходами только тех испытаний, которые могут быть воспроизведены неограниченное число раз при одном и том же комплексе условий.* Так, например, бессмысленно ставить вопрос об определении вероятностей возникновения войн, появления гениальных произведений искусства и т.п., так как речь идет о неповторимых в одинаковых условиях испытаниях, уникальных событиях. Или, например, не имеет смысла говорить о том, что данный студент сдаст семестровый экзамен по теории вероятностей, поскольку речь здесь идет о единичном испытании, повторить которое в тех же условиях нет возможности.

И хотя приведенные в примерах события с неопределенным исходом относятся к категории «может произойти, а может и не произойти», такими событиями теория вероятностей не занимается.

2. События должны обладать так называемой *статистической устойчивостью*, или *устойчивостью относительных частот.* Это означает, что в различных сериях испытаний относительная частота (частость) события изменяется незначительно (тем

меньше, чем больше число испытаний), колеблясь около постоянного числа. Оказалось, что этим постоянным числом является вероятность события (об этом идет речь в теореме Бернулли, приведенной в гл. 6).

Факт приближения относительной частоты, или частости, события к его вероятности при увеличении числа испытаний, сводящихся к схеме случаев, подтверждается многочисленными массовыми экспериментами, проводимыми разными лицами со времен возникновения теории вероятностей. Так, например, в опытах Бюффона (XVIII в.) относительная частота (частость) появления герба при 4040 подбрасываниях монеты оказалась равной 0,5069, в опытах Пирсона (XIX в.) при 23000 подбрасываниях — 0,5005, практически не отличаясь от вероятности этого события, равной 0,5.

3. *Число испытаний*, в результате которых появляется событие A, *должно быть достаточно велико*, ибо только в этом случае можно считать вероятность события $P(A)$ приближенно равной ее относительной частоте.

Резюмируя, можно сказать, что *теория вероятностей изучает лишь такие события, в отношении которых имеет смысл не только утверждение об их случайности, но и возможна объективная оценка относительной частоты их появления.* Так, утверждение, что при выполнении определенного комплекса условий S вероятность события равна p, означает не только *случайность события* A, но и *определенную*, достаточно близкую к p, *долю появлений события* A *при большом числе испытаний*; а значит, выражает *определенную объективную* (хотя и своеобразную) *связь между комплексом условий* S *и событием* A (не зависящую от субъективных суждений о наличии этой связи того или иного лица). И даже просто существование вероятности p (когда само значение p неизвестно) сохраняет качественно суть этого утверждения, выделенную курсивом.

Легко проверить, что свойства вероятности (см. **(1.2)**), вытекающие из классического определения **(1.1)**, сохраняются и при статистическом определении вероятности **(1.3)**.

1.4. Геометрическое определение вероятности

Одним из недостатков классического определения вероятности **(1.1)**, ограничивающим его применение, является то, что

оно предполагает конечное число возможных исходов испытания.

Оказывается, иногда этот недостаток можно преодолеть, используя геометрическое определение вероятности, т.е. находя вероятность попадания точки в некоторую область (отрезок, часть плоскости и т.п.).

Пусть, например, плоская фигура g составляет часть плоской фигуры G. На фигуру G наудачу бросается точка. Это означает, что все точки области G «равноправны» в отношении попадания туда брошенной случайной точки. Полагая, что вероятность события A — попадания брошенной точки на фигуру g — пропорциональна площади этой фигуры и не зависит ни от ее расположения относительно G, ни от формы g, найдем

$$P(A)=\frac{S_g}{S_G}, \qquad \textbf{(1.4)}$$

где S_g и S_G — соответственно площади областей g и G (рис. 1.1).

Фигуру g называют *благоприятствующей* (*благоприятной*) *событию* A.

Область, на которую распространяется понятие геометрической вероятности, может быть одномерной (прямая, отрезок) и трехмерной (некоторое тело в пространстве). Обозначая меру (длину, площадь, объем) области через mes, приходим к следующему определению.

Определение. ***Геометрической вероятностью*** *события* A *называется отношение меры области, благоприятствующей появлению события* A, *к мере всей области*, т.е.

$$P(A)=\frac{\operatorname{mes} g}{\operatorname{mes} G}. \qquad \textbf{(1.5)}$$

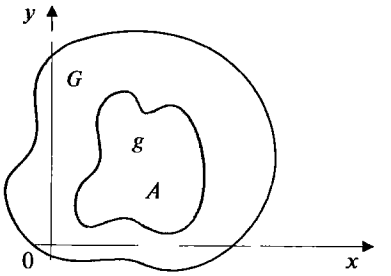


Рис. 1.1

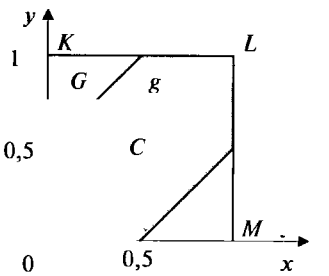


Рис. 1.2

▷ **Пример 1.2.** Два лица — A и B условились встретиться в определенном месте, договорившись только о том, что каждый является туда в любой момент времени между 11 и 12 ч и ждет в течение 30 мин. Если партнер к этому времени еще не пришел или уже успел покинуть установленное место, встреча не состоится. Найти вероятность того, что встреча состоится.

Р е ш е н и е. Обозначим моменты прихода в определенное место лиц A и B соответственно через x и y. В прямоугольной системе координат Oxy возьмем за начало отсчета 11 ч, а за единицу измерения — 1 ч. По условию $0 \le x \le 1$, $0 \le y \le 1$. Этим неравенствам удовлетворяют координаты любой точки, принадлежащей квадрату $OKLM$ со стороной, равной 1 (рис. 1.2). Событие C — встреча двух лиц— произойдет, если разность между x и y не превзойдет 0,5 ч (по абсолютной величине), т.е. $|y - x| \le 0{,}5$.

Решение последнего неравенства есть полоса $x-0{,}5 \le y \le x+0{,}5$, которая внутри квадрата на рис. 1.2 представляет заштрихованную область g. По формуле **(1.4)**

$$P(C) = \frac{S_g}{S_G} = \frac{1 - 2(1\ 2)\cdot 0{,}5^2}{1^2} = 0{,}75,$$

так как площадь области g равна площади квадрата G без суммы площадей двух угловых (незаштрихованных) треугольников. ▶

1.5. Элементы комбинаторики

Для успешного решения задач с использованием классического определения вероятности необходимо знать основные правила и формулы *комбинаторики* — раздела математики, изучающего, в частности, методы решения *комбинаторных задач* — задач на подсчет числа различных комбинаций.

Пусть A_i (i = 1,2,...,n) — элементы конечного множества. Сформулируем два важных правила, часто применяемых при решении комбинаторных задач.

Правило суммы. *Если элемент A_1 может быть выбран n_1 способами, элемент A_2 — другими n_2 способами, A_3 — отличными от первых двух n_3 способами и т.д., A_k — n_k способами, отличными от первых (k–1), то выбор одного из элементов: или A_1, или A_2,..., или A_k может быть осуществлен $n_1+n_2+...+n_k$ способами.*

▷ **Пример 1.3.** В ящике 300 деталей. Известно, что 150 из них — 1-го сорта, 120 — 2-го, а остальные — 3-го сорта. Сколько существует способов извлечения из ящика одной детали 1-го или 2-го сорта?

Р е ш е н и е. Деталь 1-го сорта может быть извлечена n_1=150 способами, 2-го сорта — n_2=120 способами. По правилу суммы существует n_1+n_2 = 150+120=270 способов извлечения одной детали 1-го или 2-го сорта. ▶

Правило произведения. *Если элемент A_1 может быть выбран n_1 способами, после каждого такого выбора элемент A_2 может быть выбран n_2 способами и т.д., после каждого (k–1) выбора элемент A_k может быть выбран n_k способами, то выбор всех элементов A_1, A_2,...,A_k в указанном порядке может быть осуществлен $n_1n_2...n_k$ способами.*

▷ **Пример 1.4.** В группе 30 человек. Необходимо выбрать старосту, его заместителя и профорга. Сколько существует способов это сделать?

Р е ш е н и е. Старостой может быть выбран любой из 30 учащихся, его заместителем — любой из оставшихся 29, а профоргом — любой из оставшихся 28 учащихся, т.е. n_1=30, n_2 =29, n_3=28. По правилу произведения общее число способов выбора старосты, его заместителя и профорга равно $n_1n_2n_3 = 30\cdot 29\cdot 28 =$ = 24 360 способов. ▶

Пусть дано множество из n различных элементов. Из этого множества могут быть образованы подмножества из m элементов ($0 \le m \le n$). Например, из 5 элементов a, b, c, d, e могут быть отобраны комбинации по 2 элемента — ab, cd, eb, ba, ce и т.д., по 3 элемента — abc, cbd, cba, ead и т.д.

Если комбинации из n элементов по m отличаются либо составом элементов, либо порядком их расположения (либо и тем и другим), то такие комбинации называют *размещениями* из n элементов по m. Число размещений из n элементов по m равно

$$A_n^m = \underbrace{n(n-1)(n-2)\ldots(n-m+1)}_{m \text{ сомножителей}} \qquad \textbf{(1.6)}$$

или

$$A_n^m = \frac{n!}{(n-m)!}, \qquad \textbf{(1.7)}$$

где $n!$ равно произведению n первых чисел натурального ряда, т.е. $n! = 1\cdot 2...n$.

▷ **Пример 1.5.** Расписание одного дня состоит из 5 уроков. Определить число вариантов расписания при выборе из 11 дисциплин.

Р е ш е н и е. Каждый вариант расписания представляет набор 5 дисциплин из 11, отличающийся от других вариантов как составом дисциплин, так и порядком их следования (или и тем и другим), т.е. является размещением из 11 элементов по 5. Число вариантов расписаний, т.е. число размещений из 11 по 5, находим по формуле **(1.6)**

$$A_{11}^5 = \underbrace{11\cdot 10\cdot 9\cdot 8\cdot 7}_{5-\text{сомножителей}} = 55\,440. \blacktriangleright$$

Если комбинации из n элементов по m отличаются только составом элементов, то их называют *сочетаниями* из n элементов по m. Число сочетаний из n элементов по m равно

$$C_n^m = \frac{n(n-1)(n-2)\ldots(n-m+1)}{1\cdot 2\ldots m} \qquad \textbf{(1.8)}$$

или

$$C_n^m = \frac{n!}{m!(n-m)!}. \qquad \textbf{(1.9)}$$

Так как по определению $0! = 1$, то $C_n^0 = 1$.

Свойства числа сочетаний:

$$C_n^m = C_n^{n-m}, \qquad \textbf{(1.10)}$$

$$C_n^m + C_n^{m+1} = C_{n+1}^{m+1} \qquad \textbf{(1.11)}$$

▷ **Пример 1.6.** В шахматном турнире участвуют 16 человек Сколько партий должно быть сыграно в турнире, если между любыми двумя участниками должна быть сыграна одна партия?

Р е ш е н и е. Каждая партия играется двумя участниками из 16 и отличается от других только составом пар участников, т.е. представляет собой сочетание из 16 элементов по 2. Их число находим по формуле **(1.8)**:

$$C_{16}^2 = \frac{16\cdot 15}{1\cdot 2} = 120. \blacktriangleright$$

Если комбинации из n элементов отличаются только порядком расположения этих элементов, то их называют *перестановками* из n элементов. Число перестановок из n элементов равно

$$P_n = n! \qquad \textbf{(1.12)}$$

▷ **Пример 1.7.** Порядок выступления 7 участников конкурса определяется жребием. Сколько различных вариантов жеребьевки при этом возможно?

Р е ш е н и е. Каждый вариант жеребьевки отличается только порядком участников конкурса, т.е. является перестановкой из 7 элементов. Их число по формуле **(1.12)**: $P_7 = 7! = 1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7 = 5040.$ ▶

Если в размещениях (сочетаниях) из n элементов по m некоторые из элементов (или все) могут оказаться одинаковыми, то такие размещения (сочетания) называют *размещениями* (*сочетаниями*) *с повторениями из n* элементов по m.

Например, из 5 элементов a, b, c, d, e по 3 размещениями с повторениями будут *abc*, *cba*, *bcd*, *cdb*, *bbe*, *ebb*, *beb*, *ddd* и т.д., сочетаниями с повторениями будут *abc*, *bcd*, *bbe*, *ddd* и т.д.

Число размещений с повторениями из n элементов по m равно

$$\tilde{A}_n^m = n^m, \qquad \textbf{(1.13)}$$

а число сочетаний с повторениями из n элементов по m равно

$$\tilde{C}_n^m = C_{n+m-1}^m, \qquad \textbf{(1.14)}$$

где C_{n+m-1}^m определяется по формуле **(1.8)** или **(1.9)**.

▷ **Пример 1.8.** В конкурсе по 5 номинациям участвуют 10 кинофильмов. Сколько существует вариантов распределения призов, если по каждой номинации установлены: а) различные призы; б) одинаковые призы?

Р е ш е н и е. а) Каждый из вариантов распределения призов представляет собой комбинацию 5 фильмов из 10, отличающуюся от других комбинаций как составом фильмов, так и их порядком по номинациям (или и тем и другим), причем одни и те же фильмы могут повторяться несколько раз[1], т.е. представляет размещение с повторениями из 10 элементов по 5. Их число по формуле **(1.13)** равно

$$\tilde{A}_{10}^5 = 10^5 = 100\,000.$$

[1]Любой фильм может получить призы как по одной, так и по нескольким (включая все пять) номинациям.

б) Если по каждой номинации установлены одинаковые призы, то порядок следования фильмов в комбинации 5 призеров значения не имеет, и число вариантов распределения призов представляет собой число сочетаний с повторениями из 10 элементов по 5, определяемое по формуле (**1.14**) с учетом (**1.8**):

$$\widetilde{C}_{10}^{5} = C_{10+5-1}^{5} = C_{14}^{5} = \frac{14 \cdot 13 \cdot 12 \cdot 11 \cdot 10}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 2002. \blacktriangleright$$

Если в перестановках из общего числа n элементов есть k различных элементов, при этом 1-й элемент повторяется n_1 раз, 2-й элемент — n_2 раз, k-й элемент — n_k раз, причем $n_1+n_2+...+n_k=n$, то такие перестановки называют *перестановками с повторениями* из n элементов. Число перестановок с повторениями из n элементов равно

$$P_n(n_1, n_2, ..., n_k) = \frac{n!}{n_1! n_2! ... n_k!}. \qquad (1.15)$$

▷ **Пример 1.9.** Сколько существует семизначных чисел, состоящих из цифр 4, 5 и 6, в которых цифра 4 повторяется 3 раза, а цифры 5 и 6 — по 2 раза?

Р е ш е н и е. Каждое семизначное число отличается от другого порядком следования цифр (причем n_1=3, n_2=2, n_3=2, а их сумма равна 7), т.е. является перестановкой с повторениями из 7 элементов. Их число по формуле (**1.15**):

$$P_7(3;2;2) = \frac{7!}{3!\,2!\,2!} = 210. \blacktriangleright$$

1.6. Непосредственное вычисление вероятностей

Для непосредственного вычисления вероятности используется ее классическое определение (**1.1**).

▷ **Пример 1.10.** Буквы Т, Е, И, Я, Р, О написаны на отдельных карточках. Ребенок берет карточки в случайном порядке и прикладывает одну к другой: а) 3 карточки; б) все 6 карточек. Какова вероятность того, что получится слово: а) «ТОР»; б) «ТЕОРИЯ»?

Р е ш е н и е. Пусть событие A — получение слова «ТОР». Различные комбинации трех букв из имеющихся шести представляют размещения, так как могут отличаться как составом

входящих букв, так и порядком их следования (или и тем и другим), т.е. общее число случаев $n = A_6^3$, из которых благоприятствует событию A $m = 1$ случай. По формуле (**1.1**)

$$P(A) = \frac{m}{n} = \frac{1}{A_6^3} = \frac{1}{6 \cdot 5 \cdot 4} = \frac{1}{120}.$$

б) Пусть событие B — получение слова «ТЕОРИЯ». Различные комбинации шести букв из имеющихся шести представляют собой перестановки, так как отличаются только порядком следования букв; т.е. общее число случаев $n = P_6 = 6!$, из которых благоприятствует событию B $m = 1$ случай. Поэтому

$$P(B) = \frac{m}{n} = \frac{1}{P_6} = \frac{1}{6!} = \frac{1}{720}. \blacktriangleright$$

▷ **Пример 1.11.** Используя условие примера **1.10**, найти вероятность того, что получится слово «АНАНАС», если на отдельных карточках написаны три буквы А, две буквы Н и одна буква С.

Р е ш е н и е. Пусть событие B — получение слова «АНАНАС». Так же, как и в примере **1.10** б, общее число случаев $n = P_6 = 6!$, но теперь число случаев m, благоприятствующих событию B, существенно больше, так как перестановка трех букв А, осуществляемая $P_3 = 3!$ способами, и перестановка двух букв Н ($P_2 = 2!$ способами) не меняет собранное из карточек слово «АНАНАС»; по правилу произведения (см. § 1.5) $m = P_3 \cdot P_2$

Итак,

$$P(B) = \frac{m}{n} = \frac{P_3 \cdot P_2}{P_6} = \frac{3!\,2!}{6!} = \frac{1}{60}.$$

(Задачу можно решить и иначе, рассматривая комбинации букв как перестановки с повторениями (см. § 1.5), из которых событию B благоприятствует 1 комбинация:

$$P(B) = 1 : P_6(3;2;1) = 1 : \frac{6!}{3!\,2!\,1!} = \frac{1}{60}). \blacktriangleright$$

▷ **Пример 1.12.** Из 30 студентов 10 имеют спортивные разряды. Какова вероятность того, что выбранные наудачу 3 студента — разрядники?

Р е ш е н и е. Пусть событие A — 3 выбранных наудачу студента — разрядники. Общее число случаев выбора 3 студентов

из 30 равно $n = C_{30}^3$, так как комбинации из 30 студентов по 3 представляют собой сочетания, ибо отличаются только составом студентов. Точно так же число случаев, благоприятствующих событию A, равно $n = C_{10}^3$. Итак,

$$P(A) = \frac{m}{n} = \frac{C_{10}^3}{C_{30}^3} = \frac{10 \cdot 9 \cdot 8}{1 \cdot 2 \cdot 3} : \frac{30 \cdot 29 \cdot 28}{1 \cdot 2 \cdot 3} = \frac{61}{203} \approx 0{,}030. \blacktriangleright$$

▷ **Пример 1.13.** В лифт на 1-м этаже девятиэтажного дома вошли 4 человека, каждый из которых может выйти независимо друг от друга на любом этаже с 2-го по 9-й. Какова вероятность того, что все пассажиры выйдут: а) на 6-м этаже; б) на одном этаже?

Р е ш е н и е. а) Пусть событие A — все пассажиры выйдут на 6-м этаже. Каждый пассажир может выйти со 2-го по 9-й этаж 8 способами. По правилу произведения общее число способов выхода четырех пассажиров из лифта равно $n = 8 \cdot 8 \cdot 8 \cdot 8 = 8^4$. Число случаев, благоприятствующих событию A, равно $m = 1$. Таким образом,

$$P(A) = \frac{m}{n} = \frac{1}{8^4} = 0{,}00024.$$

б) Пусть событие B — все пассажиры выйдут на одном этаже. Теперь событию B будут благоприятствовать $m = 8$ случаев (все пассажиры выйдут или на 2-м этаже, или на 3-м,..., или на 9-м этаже). Поэтому

$$P(B) = \frac{m}{n} = \frac{8}{8^4} = \frac{1}{8^3} = 0{,}00195.$$

(Общее число способов выхода пассажиров из лифта можно найти иначе, если учесть, что комбинации номеров этажей, на которых может выйти из лифта каждый из четырех пассажиров, например, 3456, 4356, 4433, 5666, 5555, 9785 и т.д., представляют собой размещения с повторениями из 8 элементов (этажей) по 4. Их число по формуле (**1.13**) равно $n = \tilde{A}_8^4 = 8^4$.) ▶

▷ **Пример 1.14.** По условиям лотереи «Спортлото 6 из 45» участник лотереи, угадавший 4, 5, 6 видов спорта из отобранных при случайном розыгрыше 6 видов спорта из 45, получает денежный приз. Найти вероятность того, что будут угаданы: а) все 6 цифр; б) 4 цифры.

Р е ш е н и е. а) Пусть событие A — угадывание всех 6 видов спорта из 45. Общее число всех случаев, т.е. всех вариантов заполнения карточек спортлото, есть $n = C_{45}^6$, так как каждый вариант заполнения отличается только составом видов спорта. Число случаев, благоприятствующих событию A, есть $m = 1$. Поэтому

$$P(A) = \frac{1}{C_{45}^6} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6}{45 \cdot 44 \cdot 43 \cdot 42 \cdot 41 \cdot 40} \approx 0{,}0000001.$$

б) Пусть событие B — угадывание 4 видов спорта из 6 выигравших из 45. Вначале найдем число способов, какими можно выбрать 4 вида спорта из 6 выигравших, т.е. C_6^4. Но это еще не все: к каждой комбинации 4-х выигравших видов спорта из 6 следует присоединить комбинацию 2-х невыигравших видов из $45 - 6 = 39$; таких комбинаций C_{39}^2. По правилу произведения общее число случаев, благоприятствующих событию B, равно $m = C_6^4 \cdot C_{39}^2$. Итак,

$$P(B) = \frac{m}{n} = \frac{C_6^4 \cdot C_{39}^2}{C_{45}^6} = 0{,}00136. \blacktriangleright$$

▷ **Пример 1.15.** В партии 100 изделий, из которых 4 — бракованные. Партия произвольно разделена на две равные части, которые отправлены двум потребителям. Какова вероятность того, что все бракованные изделия достанутся: а) одному потребителю: б) обоим потребителям поровну?

Р е ш е н и е. а) Пусть событие A — все бракованные изделия достанутся одному потребителю. Общее число способов, какими можно выбрать 50 изделий из 100, равно $n = C_{100}^{50}$. Событию A благоприятствуют случаи, когда из 50 изделий, отправленных одному потребителю, будет либо 46 стандартных из 96 (и все 4 бракованных) изделий, либо 50 стандартных из 96 (и 0 бракованных); их число $m = C_{96}^{46} \cdot C_4^4 + C_{96}^{50} C_4^0$. Поэтому

$$P(A) = \frac{m}{n} = \frac{C_{96}^{46} \cdot C_4^4 + C_{96}^{50} \cdot C_4^0}{C_{100}^{50}} = \frac{C_{96}^{46} \cdot 1 + C_{96}^{46} \cdot 1}{C_{100}^{50}} = \frac{2C_{96}^{46}}{C_{100}^{50}} =$$

$$= \frac{2 \cdot 96! \cdot 50! \cdot 50!}{46! \cdot 50! \cdot 100!} = \frac{2 \cdot 96! \cdot 46! \cdot 47 \cdot 48 \cdot 49 \cdot 50}{46! \cdot 96! \cdot 97 \cdot 98 \cdot 99 \cdot 100} = 0{,}117,$$

где $100! = 96! \cdot 97 \cdot 98 \cdot 99 \cdot 100$, $50! = 46! \cdot 47 \cdot 48 \cdot 49 \cdot 50$.

б) Пусть событие B — в каждой партии по 2 бракованных изделия. Теперь событию B будут благоприятствовать случаи, когда из 50 изделий, отправленных одному потребителю, будут 48 стандартных из 96 и 2 бракованных из 4, их число $m = C_{96}^{48} \cdot C_4^2$. Поэтому

$$P(B) = \frac{m}{n} = \frac{C_{96}^{48} \cdot C_4^2}{C_{100}^{50}} = \frac{96! \cdot 4! \cdot 50! \cdot 50!}{48! \cdot 48! \cdot 2! \cdot 2! \cdot 100!} =$$

$$\frac{96!(2! \cdot 3 \cdot 4)(48! \cdot 49 \cdot 50)^2}{(48!)^2 2! \cdot 2(96! \cdot 97 \cdot 98 \cdot 99 \cdot 100)} = \frac{3 \cdot 4(49 \cdot 50)^2}{2 \cdot 97 \cdot 98 \cdot 99 \cdot 100} = 0{,}383. \blacktriangleright$$

▷**Пример 1.16.** В магазине было продано 21 из 25 холодильников трех марок, имеющихся в количествах 5, 7 и 13 штук. Полагая, что вероятность быть проданным для холодильника каждой марки одна и та же, найти вероятность того, что остались нераспроданными холодильники: а) одной марки; б) трех разных марок.

Р е ш е н и е. а) Пусть событие A — остались нераспроданными холодильники одной марки. Общее число способов, которыми можно получить 4 (непроданных) холодильника из 25, равно $n = C_{25}^4$. Число способов, которыми можно получить 4 холодильника первой марки из 5, равно $m_1 = C_5^4$; второй марки из 7 — $m_2 = C_7^4$ и третьей марки из 13 — $m_3 = C_{13}^4$. Событию A по правилу суммы (§ 1.5) благоприятствует $m = m_1 + m_2 + m_3 = C_5^4 + C_7^4 + C_{13}^4$ случаев. Поэтому

$$P(A) = \frac{m}{n} = \frac{C_5^4 + C_7^4 + C_{13}^4}{C_{25}^4} = \frac{5 + 35 + 715}{12\,650} = \frac{755}{12\,650} = 0{,}060.$$

б) Пусть событие B — остались нераспроданными холодильники трех разных марок. Событие B может произойти по одному из трех вариантов. По первому варианту событие B произойдет, если останутся нераспроданными 1, 1, 2 холодильников соответственно 1-й, 2-й и 3-й марок; по второму варианту — 1, 2, 1 и по третьему варианту останутся нераспроданными 2, 1, 1 холодильников соответственно 1-й, 2-й и 3-й марок. Так как до продажи имелось 5 холодильников 1-й марки, 7 — 2-й и 13 холодильников 3-й марки, то по правилу произведения (§ 1.5) число случаев, благоприятствующих первому варианту, равно $m_1 = C_5^1 C_7^1 C_{13}^2$; второму — $m_2 = C_5^1 C_7^2 C_{13}^1$; третьему варианту — $m_3 = C_5^2 C_7^1 C_{13}^1$. Общее число случаев, благоприятствующих событию B, равно $m = m_1 + m_2 + m_3$. Теперь

$$P(B) = \frac{m}{n} = \frac{m_1 + m_2 + m_3}{n} = \frac{C_5^1 C_7^1 C_{13}^2 + C_5^1 C_7^2 C_{13}^1 + C_5^2 C_7^1 C_{13}^1}{C_{25}^4} =$$

$$= \frac{5 \cdot 7 \cdot 78 + 5 \cdot 21 \cdot 13 + 10 \cdot 7 \cdot 13}{12\,650} = \frac{5005}{12\,650} = 0{,}396. \blacktriangleright$$

▷**Пример 1.16а.** В аудитории $m = 25$ студентов. Найти вероятность того, что хотя бы у двух студентов дни рождения совпадают. При каком числе m студентов вероятность того же события не меньше чем 0,95? (Полагаем равновозможность рождений в любой день года.)

Р е ш е н и е. Пусть событие A — дни рождения хотя бы двух студентов из m присутствующих в аудитории совпадают. Найдем вероятность противоположного события $\bar{A}$ — дни рождения всех студентов различны.

Число случаев, благоприятствующих событию A, есть число размещений из $n = 365$ элементов (дней года) по m, т.е. A_n^m. Общее число случаев определяется также числом размещений из n элементов по m, но размещений с повторениями, т.е. $\tilde{A}_n^m = n^m$ (см. **(1.13)**). Согласно классическому определению вероятности

$$P(\bar{A}) = \frac{A_n^m}{n^m} = \frac{n(n-1)...(n-m+1)}{n^m}$$

и для $n = 365$

$$P(A) = 1 - P(\bar{A}) = 1 - \frac{365 \cdot 364...(365-m+1)}{365^m} \qquad (*)$$

При $m = 25$ искомая вероятность, рассчитанная по формуле (*), составит $P(A) = 0{,}569$.

Вычисляя вероятности $P(A)$ для различных m, нетрудно убедиться в том, что неравенство $P(A) > 0{,}95$ будет выполняться при $m \geq 47$, т.е. достаточно лишь 47 студентов в аудитории, чтобы с вероятностью, не меньшей чем 0,95, утверждать, что по крайней мере у двух из них дни рождения совпадают. ▶

1.7. Действия над событиями

Введем понятие суммы, произведения и разности событий.

Определение. ***Суммой*** *нескольких событий называется событие, состоящее в наступлении хотя бы одного из данных событий.*

Если A и B — совместные события, то их сумма[1] $A + B$ обозначает наступление или события A, или события B, или обоих событий вместе. Если A и B — несовместные события, то их сумма $A + B$ означает наступление или события A, или события B.

Определение. ***Произведением*** *нескольких событий называется событие, состоящее в совместном наступлении всех этих событий.*

Если A, B, C — совместные события, то их произведение[1] ABC означает наступление и события A, и события B, и события C.

Определение. ***Разностью***[1] *A — B двух событий A и B называется событие, которое состоится, если событие A произойдет, а событие B не произойдет.*

▷ **Пример 1.17.** Победитель соревнования награждается: призом (событие A), денежной премией (событие B), медалью (событие C). Что представляют собой события: а) $A + B$; б) ABC; в) $AC - B$?

Решение. а) Событие $A + B$ состоит в награждении победителя или призом, или премией, или и тем и другим.

б) Событие ABC состоит в награждении победителя одновременно и призом, и премией, и медалью.

в) Событие $AB - C$ состоит в награждении победителя одновременно и призом, и премией без выдачи медали. ▶

Ниже (§ 1.12) рассматривается теоретико-множественная трактовка основных понятий теории вероятностей. Здесь же дадим геометрическую интерпретацию основных действий над событиями с помощью *диаграмм Венна.*

Пусть, например, внутри прямоугольника (рис. 1.3) выбирается наудачу точка (достоверное событие Ω), и событие A состоит в попадании этой точки в меньший круг (рис. 1.3*а*), а событие B — в больший круг (рис. 1.3*б*). Тогда сумма событий $A + B$ означает попадание точки во всю заштрихованную область обоих кругов (рис. 1.3*в*), а произведение AB — в общую часть кругов (рис. 1.3*г*). На рис. 1.3*д*, *е* заштрихованные области показывают события $\overline{A}$ и $\overline{B}$, противоположные событиям A и B, а на рис. 1.3*ж* и *з* — разности событий $A - B$ и $B - A$.

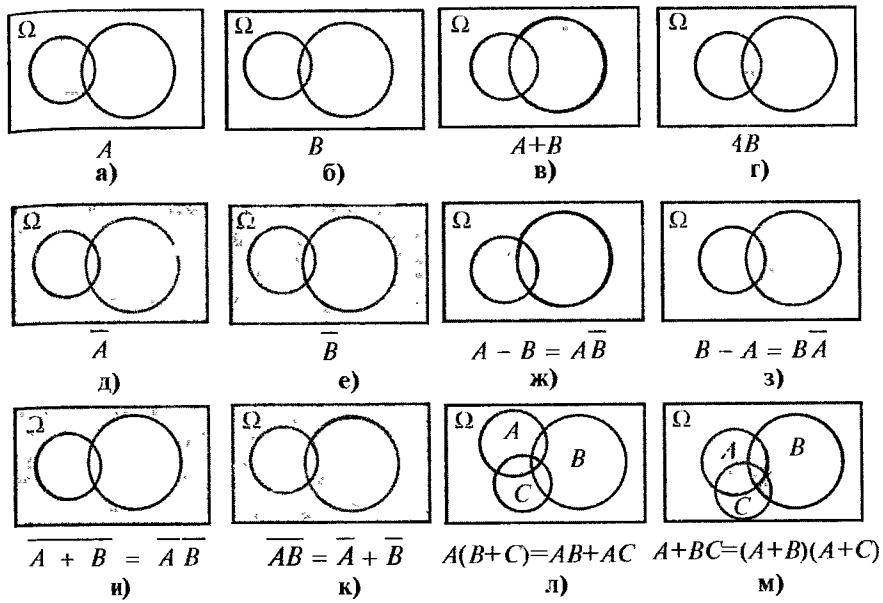


Рис. 1.3

▷ **Пример 1.18.** Убедиться в справедливости равенств:

а) $\overline{A+B+...+K} = \overline{A}\,\overline{B}...\overline{K}$; б) $\overline{AB...K} = \overline{A} + \overline{B} + ... + \overline{K}$.

Решение. а) Если событие $A+B+...+K$ состоит в появлении хотя бы одного из данных событий A, B,..., K, то противоположное событие $\overline{A+B+...+K}$ означает непоявление всех данных событий, т.е. произведение событий $\overline{A}\,\overline{B}...\overline{K}$.

б) Если событие $AB...K$ состоит в совместном наступлении всех данных событий A, B,..., K, то противоположное событие $\overline{AB...K}$ означает непоявление хотя бы одного из этих событий, т.е. сумму $\overline{A} + \overline{B} + ... + \overline{K}$. На рис. 1.3*и* и *к* приведенные соотношения между событиями иллюстрируются на примере двух событий. ▶

Если событие A представляет собой сумму несовместных событий $A_1, A_2, ..., A_n$, т.е. $A = A_1 + A_2 + ... + A_n$, то говорят, что событие A распадается на *n частных случаев* (*вариантов*) $A_1, A_2, ..., A_n$.

[1] Для суммы событий A и B используется также обозначение $A \cup B$, для произведения тех же событий — $A \cap B$, а для их разности — $A \setminus B$ (см. § 1.12).

Операции сложения и умножения событий обладают следующими свойствами:

1. $A + B = B + A$ — коммутативность сложения.

2. $A + (B + C) = (A + B) + C$ — ассоциативность сложения.

3. $AB = BA$ — коммутативность умножения.

4. $A(BC) = (AB)C$ — ассоциативность умножения.

5. $A(B + C) = AB + AC$; $A + BC = (A + B)(A + C)$ — законы дистрибутивности.

Последние два свойства иллюстрируются на рис. 1.3*л* и *м*.

Из определения операций над событиями вытекают очевидные равенства:

$$A+A=A,\ AA=A;\ A+\Omega=\Omega,\ A\Omega=A;\ A+\varnothing=A,\ A\varnothing=\varnothing.$$

1.8. Теорема сложения вероятностей

Сформулируем теорему (правило) сложения вероятностей.

Теорема. *Вероятность суммы конечного числа несовместных событий равна сумме вероятностей этих событий:*

$$P(A + B + \ldots + K) = P(A) + P(B) + \ldots + P(K). \quad \textbf{(1.16)}$$

□ Докажем теорему для схемы случаев, рассматривая сумму двух событий.

Пусть в результате испытания из общего числа n равновозможных и несовместных (элементарных) исходов испытания (случаев) событию A благоприятствует m_1 случаев, а событию B — m_2 случаев (рис. 1.4).

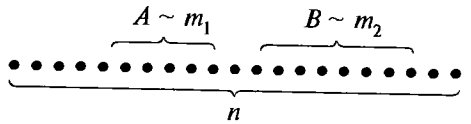


Рис. 1.4

Согласно классическому определению $P(A) = \frac{m_1}{n}$, $P(B) = \frac{m_2}{n}$.

Так как события A и B несовместные, то ни один из случаев, благоприятствующих одному из этих событий, не благоприятствует другому (см. рис. 1.4). Поэтому событию $A+B$ будет благоприятствовать m_1+m_2 случаев. Следовательно,

$$P(A + B) = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n} = P(A) + P(B)$$ ■

Следствие 1. *Сумма вероятностей событий, образующих полную группу, равна единице:*

$$P(A) + P(B) + \ldots + P(K) = 1. \quad \textbf{(1.17)}$$

□ Если события $A,B,\ldots,K$ образуют полную группу, то они единственно возможные и несовместные.

Так как события $A,B,\ldots,K$ — единственно возможные, то событие $A+B+\ldots+K$, состоящее в появлении в результате испытания хотя бы одного из этих событий, является достоверным[1], т е. его вероятность равна единице:

$$P(A + B + \ldots + K) = 1.$$

В силу того, что события $A,B,\ldots,K$ — несовместные, к ним применима теорема сложения **(1.16)**, т.е.

$$P(A + B + \ldots + K) = P(A) + P(B) + \ldots + P(K) = 1.$$ ■

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

$$P(A) + P(\overline{A}) = 1. \quad \textbf{(1.18)}$$

□ Утверждение **(1.18)** следует из того, что противоположные события образуют полную группу. ■

▷ **Пример 1.19.** Вероятность выхода изделия из строя при эксплуатации сроком до одного года равна 0,13, а при эксплуатации сроком до 3 лет — 0,36. Найти вероятность выхода изделия из строя при эксплуатации сроком от 1 года до 3 лет.

Решение. Пусть события A, B, C — выход из строя изделий при эксплуатации сроком соответственно до 1 года, от 1 года до 3 лет, свыше 3 лет, причем по условию $P(A) = 0{,}13$, $P(C) = 0{,}36$. Очевидно, что $C = A + B$, где A и B — несовместные события. По теореме сложения $P(C) = P(A) + P(B)$, откуда $P(B) = P(C) - P(A) = 0{,}36 - 0{,}13 = 0{,}23$. ◀

Замечание. Следует еще раз подчеркнуть, что рассмотренная *теорема сложения применима только для несовместных*

[1] Поэтому полную группу событий можно было бы определить и иначе, чем в § 1.1: несколько событий образуют *полную группу* (*систему*), если они являются несовместными исходами испытания и их сумма представляет собой достоверное событие.

событий и попытка ее использования в виде **(1.16)** для совместных событий приводит к неверным и даже абсурдным результатам. Например, пусть вероятность события A_i — выигрыша по любому билету денежно-вещевой лотереи, т.е. $P(A_i) = 0{,}05$, и приобретено 100 билетов ($i = 1,2,...,100$). Тогда, применяя теорему сложения, получим, что вероятность выигрыша хотя бы по одному из 100 билетов, т.е.

$$P(A_1+A_2+...+A_i+...+A_{100}) = P(A_1)+P(A_2)+...+P(A_i)+...+P(A_{100}) =$$

$$= \underbrace{0{,}05+0{,}05+...+0{,}05}_{100\text{ раз}} = 5.$$

Абсурдность полученного ответа (вероятность любого события не может быть больше 1) объясняется неприменимостью в данном случае теоремы сложения, ибо выигрыш по каждому билету, т.е. события $A_1, A_2, ..., A_{100}$ являются событиями совместными.

1.9. Условная вероятность события. Теорема умножения вероятностей. Независимые события

Как отмечено выше, вероятность $P(B)$ как мера степени объективной возможности наступления события B имеет смысл при выполнении определенного комплекса условий. При изменении условий вероятность события B может измениться. Так, если к комплексу условий, при котором изучалась вероятность $P(B)$, добавить новое условие A, то *полученная вероятность события B, найденная при условии, что событие A произошло, называется условной вероятностью события B* и обозначается $P_A(B)$, или $P(B/A)$, или $P(B \mid A)$.

Строго говоря, «безусловная» вероятность $P(B)$ также является условной, так как она получена при выполнении определенного комплекса условий.

▷ **Пример 1.20.** В ящике 5 деталей, среди которых 3 стандартные и 2 бракованные. Поочередно из него извлекается по одной детали (с возвратом и без возврата). Найти условную вероятность извлечения во второй раз стандартной детали при условии, что в первый раз извлечена деталь: а) стандартная; б) нестандартная.

Р е ш е н и е. Пусть события A и B — извлечение стандартной детали соответственно в 1-й и 2-й раз. Очевидно, что $P(A) = \frac{3}{5}$ Если вынутая деталь вновь возвращается в ящик, то

вероятность извлечения стандартной детали во второй раз $P(B) = \frac{3}{5}$. Если вынутая деталь в ящик не возвращается то вероятность извлечения стандартной детали во второй раз $P(B)$ зависит от того, какая деталь была извлечена в первый раз — стандартная (событие A) или бракованная (событие $\overline{A}$). В первом случае $P_A(B) = \frac{2}{4}$, во втором случае $P_{\overline{A}}(B) = \frac{3}{4}$, так как из оставшихся четырех деталей стандартных будет соответственно[1] 2 или 3. ▶

Найдем формулу для вычисления условной вероятности $P_A(B)$.

□ Пусть из общего числа n равновозможных и несовместных (элементарных) исходов испытания (случаев) событию A благоприятствует m случаев, событию B — k случаев, а совместному появлению событий A и B, т.е. событию AB — l случаев ($l \le m$, $l \le k$) (рис. 1.5).

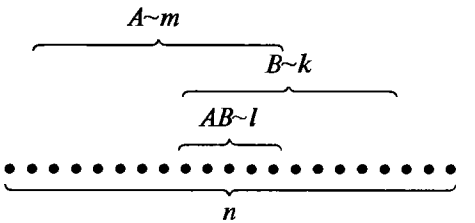


Рис. 1.5

Тогда, согласно классическому определению вероятности, $P(A) = \frac{m}{n}$, $P(AB) = \frac{l}{n}$.

После того как событие A произошло, число всех равновозможных исходов (случаев) сократилось с n до m, а число случаев, благоприятствующих событию B, с k до l. Поэтому условная вероятность[2]

[1] Следует заметить, что «безусловная» вероятность извлечения во второй раз стандартной детали $P(B)$ (когда извлеченная деталь не возвращается) определится по формуле полной вероятности **(1.31)** — см. далее § 1.11:

$P(B) = P(A)\cdot P_A(B) + P(\overline{A})P_{\overline{A}}(B) = \frac{3}{5}\cdot\frac{2}{4} + \frac{2}{5}\cdot\frac{3}{4} = \frac{3}{5}$, т.е. та же, что и при возврате извлеченной детали.

[2] Формулу условной вероятности **(1.19)** мы получили, опираясь на классическое определение вероятности. В общем случае эта формула служит определением условной вероятности (см. § 1.12).

$$P_A(B) = \frac{l}{m} = \frac{l/n}{m/n} = \frac{P(AB)}{P(A)}. \qquad (1.19)$$

Аналогично

$$P_B(A) = \frac{P(AB)}{P(B)}. \blacksquare \qquad (1.20)$$

Умножая правую и левую части равенств **(1.19)** и **(1.20)** соответственно на $P(A)$ и $P(B)$, получим

$$P(AB) = P(A) \cdot P_A(B) = P(B) \cdot P_B(A). \qquad (1.21)$$

Это так называемая **теорема (правило) умножения вероятностей**: *вероятность произведения двух событий равна произведению вероятности одного из них на условную вероятность другого, найденную в предположении, что первое событие произошло.*

Теорема (правило) умножения вероятностей[1] легко обобщается на случай произвольного числа событий:

$$P(ABC...KL) = P(A) \cdot P_A(B) \cdot P_{AB}(C)...P_{ABC...K}(L), \qquad (1.22)$$

т.е. *вероятность произведения нескольких событий равна произведению вероятности одного из этих событий на условные вероятности других; при этом условная вероятность каждого последующего события вычисляется в предположении, что все предыдущие события произошли.*

▷**Пример 1.21.** Работа электронного устройства прекратилась вследствие выхода из строя одного из пяти унифицированных блоков. Производится последовательная замена каждого блока новым до тех пор, пока устройство не начнет работать. Какова вероятность того, что придется заменить: а) 2 блока; б) 4 блока?

Р е ш е н и е. а) Обозначим события:

A_i — i-й блок исправен, $i = 1, 2, ..., 5$;

B — замена двух блоков.

Очевидно, что придется заменить 2 блока, если 1-й блок исправен (4 шанса из 5), а 2-й — неисправен (1 шанс из оставшихся 4), т.е. $B = A_1 \overline{A}_2$. Теперь по теореме умножения **(1.21)**

[1] В случае, если $P(A)=0$ или $P(B)=0$, то соответствующие формулы **(1.19)** и **(1.20)** для условных вероятностей не имеют смысла, ибо невозможно событие A или B, однако теорема (правило) умножения вероятностей **(1.21)** остается верной и при $P(A)=0$, $P(B)=0$.

$$P(B) = P(A_1 A_2) = m(A_1) \cdot m_{A_1}(A_2) = \frac{4}{5} \cdot \frac{1}{4} = \frac{1}{5}.$$

б) Пусть событие C — замена 4 блоков. Очевидно, что $C = A_1A_2A_3\overline{A}_4$ и по теореме умножения **(1.22)**

$$P(C) = P(A_1A_2A_3\overline{A}_4) = P(A_1)P_{A_1}(A_2)P_{A_1A_2}(A_3)P_{A_1A_2A_3}(\overline{A}_4) =$$

$$= \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{5}. \blacktriangleright$$

▷**Пример 1.22.** Решить другим способом задачу, приведенную в примере **1.11**.

Р е ш е н и е. Пусть событие B — получение слова «АНАНАС». Событие B наступит, если первой окажется карточка с буквой А (3 шанса из 6), вторая — с буквой Н (2 шанса из оставшихся 5), третья — с буквой А (2 шанса из оставшихся 4) и т.д. По теореме умножения

$$P(B) = \frac{3}{6} \cdot \frac{2}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{1} = \frac{1}{60}. \blacktriangleright$$

Теорема умножения вероятностей принимает наиболее простой вид, когда события, образующие произведение, *независимы*.

Событие B называется *независимым от события* A, если его вероятность не меняется от того, произошло событие A или нет, т.е.

$$P_A(B) = P(B) \quad (\text{или } P_A(B) = P(B)).$$

В противном случае, если $P_A(B) \neq P(B)$ $(\text{или } P_A(B) \neq P(B))$, событие B называется *зависимым* от A.

Докажем, что *если событие B не зависит от A, то и событие A не зависит от B.*

□ Так как по условию событие B не зависит от A, то $P_A(B) = P(B)$.

Запишем теорему умножения вероятностей **(1.21)** в двух формах:

$$P(AB) = P(A) \cdot P_A(B) = P(B) \cdot P_e(A).$$

Заменяя $P_A(B)$ на $P(B)$, получим $P(A) \cdot P(B) = P(B)P_B(A)$, откуда, полагая, что $P(B) \neq 0$, получим $P_B(A) = P(A)$, т.е. событие A не зависит от B. ■

Таким образом, *зависимость и независимость событий всегда взаимны*. Поэтому можно дать следующее определение независимости событий.

Два события называются *независимыми*, если появление одного из них не меняет вероятности наступления другого.

▷**Пример 1.23.** Установить, зависимы или нет события A и B по условию примера **1.20**.

Решение. В случае возврата извлеченной детали $P_A(B)=P_{\bar{A}}(B)=P(B)=\frac{3}{5}$, т.е. события A и B независимы. Если извлеченная из ящика деталь не возвращается, то $P_A(B)\neq P_{\bar{A}}(B)\left(\frac{2}{4}\neq\frac{3}{4}\right)$, т.е. $P_A(B)\neq P(B)$ и события A и B зависимы. ▶

Несколько событий $A,B,...,L$ называются *независимыми в совокупности* (или просто *независимыми*), если независимы любые два из них и независимы любое из данных событий и любые комбинации (произведения) остальных событий. В противном случае события $A,B,...,L$ называются *зависимыми*.

Например, три события A, B, C независимы (независимы в совокупности), если независимы события A и B, A и C, B и C, A и BC, B и AC, C и AB.

Для независимых событий теорема (правило) умножения вероятностей для двух и нескольких событий примет вид[1]:

$$P(AB)=P(A)P(B), \qquad \textbf{(1.23)}$$

$$P(ABC...KL)=P(A)P(B)...P(L), \qquad \textbf{(1.24)}$$

т.е. *вероятность произведения двух или нескольких независимых событий равна произведению вероятностей этих событий.*

▷ **Пример 1.24.** Вероятность попадания в цель для первого стрелка равна 0,8, для второго — 0,7, для третьего — 0,9. Каждый из стрелков делает по одному выстрелу. Какова вероятность того, что в мишени 3 пробоины?

Решение. Обозначим события:

A_i — попадание в цель i-го стрелка ($i=1, 2, 3$);

B — в мишени три пробоины.

[1] Формулу **(1.23)** можно было бы рассматривать и в качестве определения независимости двух событий. Для определения независимости нескольких событий (в совокупности) одной формулы **(1.24)** было бы уже недостаточно.

Очевидно, что $B=A_1A_2A_3$, причем события A_1, A_2, A_3 — независимы. По теореме умножения **(1.24)** для независимых событий

$$P(B)=P(A_1A_2A_3)=P(A_1)P(A_2)P(A_3)=0{,}8\cdot0{,}7\cdot0{,}9=0{,}504. \blacktriangleright$$

Замечание. Говоря о независимости событий, отметим следующее.

1. В *основе независимости событий лежит их физическая независимость*, означающая, что множества случайных факторов, приводящих к тому или иному исходу испытания, не пересекаются (или почти не пересекаются). Например, если в цехе имеются две установки, никак не связанные между собой по условиям производства, то простой каждой установки — события независимые. Если эти установки связаны единым технологическим циклом, то простой одной из установок зависит от состояния работы другой.

Вместе с тем, если множества случайных факторов пересекаются, то появляющиеся в результате испытания события не обязательно зависимые.

Пусть, например, рассматриваются события:

A — извлечение наудачу из колоды карты пиковой масти;

B — извлечение наудачу из колоды туза.

Необходимо выяснить, являются ли события A и B зависимыми. На первый взгляд, можно предполагать зависимость событий A и B в силу пересечения случаев, им благоприятствующих: среди карт пиковой масти есть туз, а среди тузов — карта пиковой масти. Убедимся, однако, в том, что события A и B независимы.

$$P(B)=\frac{4}{36}=\frac{1}{9}\text{ (в колоде 4 туза из 36 карт)},$$

$$P_A(B)=\frac{1}{9}\text{ (в колоде 1 туз из 9 карт пиковой масти)}.$$

Итак, $P_A(B)=P(B)$, т.е. события A и B независимы[1].

2. *Попарная независимость нескольких событий* (т.е. независимость взятых из них любых двух событий) *еще не означает их независимости в совокупности*. Убедимся в этом на примере (примере С.Н. Бернштейна).

[1] Независимость событий A и B можно показать иначе, убедившись в выполнении равенства **(1.23)**. Так как $P(AB)=1/36$ (в колоде 1 пиковый туз из 36 карт), $P(A)=9/36$ (в колоде 9 карт пиковой масти из 36), $P(B)=1/9$ (см. выше), т.е. $P(AB)=P(A)\cdot P(B)\left(\frac{1}{36}=\frac{9}{36}\cdot\frac{1}{9}\right)$, следовательно, события A и B независимы.

Предположим, что грани правильного тетраэдра (треугольной пирамиды с равными ребрами) окрашены: 1-я — в красный цвет (событие A), 2-я — в зеленый (B), 3-я — в синий (C) и 4-я — во все три цвета (событие ABC). При подбрасывании тетраэдра вероятность любой грани, на которую он упадет, в своей окраске иметь одинаковый цвет равна 1/2 (так как всего граней 4, а с соответствующей окраской 2, т.е. два шанса из четырех). Таким образом, $P(A) = 1/2$, $P(B) = 1/2$, $P(C) = 1/2$.

Точно так же можно подсчитать, что

$$P_B(A) = P_C(A) = P_A(B) = P_C(B) = P_A(C) = P_B(C) = 1/2$$

(один шанс из двух), т.е. события A, B, C попарно независимы. Если же наступили одновременно два события, например, A и B, т.е. AB, то третье событие C обязательно наступит, т.е. $P_{AB}(C) = 1$ и аналогично $P_{AC}(B) = 1$, $P_{BC}(A) = 1$; следовательно, вероятность каждого из событий A, B или C изменилась, и события A, B и C в совокупности зависимы.

При решении ряда задач требуется найти вероятность суммы двух или нескольких совместных событий, т.е. вероятность появления хотя бы одного из этих событий. Напомним, что в этом случае применять теорему сложения вероятностей в виде **(1.16)** нельзя.

Теорема. *Вероятность суммы двух совместных событий равна сумме вероятностей этих событий без вероятности их произведения, т.е.*

$$P(A + B) = P(A) + P(B) - P(AB). \qquad \textbf{(1.25)}$$

□ Представим событие $A + B$, состоящее в наступлении хотя бы одного из двух событий A и B, в виде суммы трех несовместных вариантов: $A + B = A\overline{B} + \overline{A}B + AB$. Тогда по теореме сложения

$$P(A+B) = P(A\overline{B}) + P(\overline{A}B) + P(AB). \qquad \textbf{(1.26)}$$

Учитывая, что $A = A\overline{B} + AB$, $P(A) = P(A\overline{B}) + P(AB)$, откуда $P(A\overline{B}) = P(A) - P(AB)$, и аналогично $P(\overline{A}B) = P(B) - P(AB)$, получим, подставляя найденные выражения в **(1.26)**:

$$P(A+B) = [P(A) - P(AB)] + [P(B) - P(AB)] + P(AB) =$$
$$= P(A) + P(B) - P(AB). \blacksquare$$

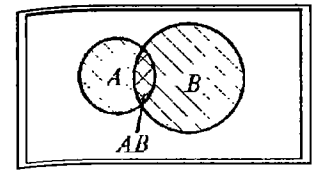


Рис. 1.6

В справедливости формулы **(1.25)** можно наглядно убедиться по рис. 1.6.

В случае трех и более совместных событий соответствующая формула для вероятности суммы $P(A+B+...+K)$ весьма громоздка и проще перейти к противоположному событию L:

$$L = \overline{A + B + ... + K} = \overline{A}\,\overline{B}...\overline{K} \text{ (см. пример } \textbf{1.18}).$$

Тогда на основании **(1.18)** $P(A+B+...+K) = 1 - P(L)$, или

$$P(A+B+...+K) = 1 - P\left(\overline{AB}...\overline{K}\right), \qquad \textbf{(1.27)}$$

т.е. *вероятность суммы нескольких совместных событий A, B, ..., K равна разности между единицей и вероятностью произведения противоположных событий $\overline{A}, \overline{B}, ..., \overline{K}$.*

Если при этом события A, B, ..., K — независимые, то

$$P(A+B+...+K) = 1 - P(\overline{A})P(\overline{B})...P(\overline{K}). \qquad \textbf{(1.28)}$$

В частном случае, если вероятности независимых событий равны, т.е. $P(A) = P(B) = ... = P(K) = p$, то вероятность их суммы

$$P(A+B+...+K) = 1 - (1-p)^n, \qquad \textbf{(1.29)}$$

(ибо в этом случае $P(\overline{A})P(\overline{B})...P(\overline{K}) = \underbrace{(1-p)...(1-p)}_{n \text{ раз}} = (1-p)^n$).

▷**Пример 1.25.** На 100 лотерейных билетов приходится 5 выигрышных. Какова вероятность выигрыша хотя бы по одному билету, если приобретено: а) 2 билета; б) 4 билета?

Р е ш е н и е. Пусть событие A_i — выигрыш по i-му билету ($i = 1, 2, 3, 4$).

а) По формуле **(1.25)** вероятность выигрыша хотя бы по одному из двух билетов

$$P(A_1 + A_2) = P(A_1) + P(A_2) - P(A_1A_2) =$$
$$= P(A_1) + P(A_2) - P(A_1)P_{A_1}(A_2) = \frac{5}{100} + \frac{5}{100} - \frac{5}{100} \cdot \frac{4}{99} = 0{,}098.$$

б) По формуле (1.27) вероятность выигрыша хотя бы по одному из четырех билетов

$$P(A_1 + A_2 + A_3 + A_4) = 1 - P(\overline{A}_1\overline{A}_2\overline{A}_3\overline{A}_4) =$$

$$= 1 - \frac{95}{100}\cdot\frac{94}{99}\cdot\frac{93}{98}\cdot\frac{92}{97} = 0{,}188. \blacktriangleright$$

1.10. Решение задач

▷ **Пример 1.26.** Вероятность того, что студент сдаст первый экзамен, равна 0,9; второй — 0,9; третий — 0,8. Найти вероятность того, что студентом будут сданы: а) только 2-й экзамен; б) только один экзамен; в) три экзамена; г) по крайней мере два экзамена; д) хотя бы один экзамен.

Р е ш е н и е. а) Обозначим события: A_i — студент сдаст i-й экзамен ($i = 1, 2, 3$); B — студент сдаст только 2-й экзамен из трех. Очевидно, что $B = \overline{A}_1 A_2 \overline{A}_3$, т.е. совместное осуществление трех событий, состоящих в том, что студент сдаст 2-й экзамен и не сдаст 1-й и 3-й экзамены. Учитывая, что события A_1, A_2, A_3 независимы, получим

$$P(B) = P(\overline{A}_1 A_2 \overline{A}_3) = P(\overline{A}_1)P(A_2)P(\overline{A}_3) = 0{,}1\cdot 0{,}9\cdot 0{,}2 = 0{,}018.$$

б) Пусть событие C — студент сдаст один экзамен из трех. Очевидно, событие C произойдет, если студент сдаст только 1-й экзамен из трех, или только 2-й, или только 3-й, т.е.

$$P(C) = P(A_1\overline{A}_2\overline{A}_3 + \overline{A}_1 A_2 \overline{A}_3 + \overline{A}_1\overline{A}_2 A_3) =$$

$$= 0{,}9\cdot 0{,}1\cdot 0{,}2 + 0{,}1\cdot 0{,}9\cdot 0{,}2 + 0{,}1\cdot 0{,}1\cdot 0{,}8 = 0{,}044.$$

в) Пусть событие D — студент сдаст все три экзамена, т.е. $D = A_1A_2A_3$. Тогда

$$P(D) = P(A_1A_2A_3) = P(A_1)P(A_2)P(A_3) = 0{,}9\cdot 0{,}9\cdot 0{,}8 = 0{,}648.$$

г) Пусть событие E — студент сдаст по крайней мере два экзамена (иначе: «хотя бы два» экзамена или «не менее двух» экзаменов). Очевидно, что событие E означает сдачу любых двух экзаменов из трех либо всех трех экзаменов, т.е.

$$E = A_1A_2\overline{A}_3 + A_1\overline{A}_2A_3 + \overline{A}_1A_2A_3 + A_1A_2A_3 \text{ и}$$

$$P(E)=0{,}9\cdot 0{,}9\cdot 0{,}2+0{,}9\cdot 0{,}1\cdot 0{,}8+0{,}1\cdot 0{,}9\cdot 0{,}8+0{,}9\cdot 0{,}9\cdot 0{,}8=0{,}954.$$

д) Пусть событие F — студент сдал хотя бы один экзамен (иначе: «не менее одного» экзамена). Очевидно, событие F представляет сумму событий C (включающего три варианта) и E (четыре варианта), т.е. $F = A_1+A_2+A_3 = C+E$ (семь вариантов). Однако проще найти вероятность события F, если перейти к противоположному событию, включающему всего один вариант — $F = \overline{A_1 + A_2 + A_3} = \overline{A}_1\overline{A}_2\overline{A}_3$, т.е. применить формулу (1.27).

Итак,

$$P(F) = P(A_1 + A_2 + A_3) = 1 - P(\overline{F}) = 1 - P(\overline{A}_1\overline{A}_2\overline{A}_3) =$$

$$= 1 - P(\overline{A}_1)\cdot P(\overline{A}_2)P(\overline{A}_3) = 1 - 0{,}1\cdot 0{,}1\cdot 0{,}2 = 0{,}998,$$

т.е. сдача хотя бы одного экзамена из трех является событием практически достоверным. ▶

▷ **Пример 1.27.** Причиной разрыва электрической цепи служит выход из строя элемента K_1 или одновременный выход из строя двух элементов — K_2 и K_3. Элементы могут выйти из строя независимо друг от друга с вероятностями, равными соответственно 0,1; 0,2; 0,3. Какова вероятность разрыва электрической цепи?

Р е ш е н и е. Обозначим события:

A_i — выход из строя элемента K_i ($i = 1, 2, 3$);

B — разрыв электрической цепи.

Очевидно, по условию событие B произойдет, если произойдет либо событие A_1, либо A_2A_3, т.е. $B = A_1 + A_2A_3$. Теперь, по формуле (1.25)

$$P(B)=P(A_1+A_2A_3)=P(A_1)+P(A_2A_3)-P[A_1(A_2A_3)]=$$

$$=P(A_1)+P(A_2)P(A_3)-P(A_1)P(A_2)P(A_3)=$$

$$=0{,}1+0{,}2\cdot 0{,}3-0{,}1\cdot 0{,}2\cdot 0{,}3=0{,}154$$

(при использовании теоремы умножения учли независимость событий A_1, A_2 и A_3). ▶

▷ **Пример 1.28.** Производительности трех станков, обрабатывающих одинаковые детали, относятся как 1:3:6. Из нерассортированной партии обработанных деталей взяты наудачу две. Какова вероятность того, что: а) одна из них обработана на 3-м станке; б) обе обработаны на одном станке?

Р е ш е н и е. а) Обозначим события:

A_i — деталь обработана на i-м станке ($i = 1, 2, 3$);

B — одна из двух взятых деталей обработана на 3-м станке.

По условию $P(A_1) = \dfrac{1}{1+3+6} = 0{,}1$, $P(A_2) = \dfrac{3}{1+3+6} = 0{,}3$, $P(A_3) = \dfrac{6}{1+3+6} = 0{,}6$.

Очевидно, что $B = A_1A_3 + A_2A_3 + A_3A_1 + A_3A_2$ (при этом надо учесть, что либо первая деталь обработана на 3-м станке, либо вторая). По теоремам сложения и умножения (для независимых событий)

$$P(B) = P(A_1)P(A_3)+P(A_2A_3)+P(A_3A_1)+P(A_3A_2)=$$

$$=0{,}1\cdot 0{,}6+0{,}3\cdot 0{,}6+0{,}6\cdot 0{,}1+0{,}6\cdot 0{,}3=0{,}48.$$

б) Пусть событие C — обе отобранные детали обработаны на одном станке. Тогда $C = A_1A_1 + A_2A_2 + A_3A_3$ и

$$P(C) = 0{,}1\cdot 0{,}1+0{,}3\cdot 0{,}3+0{,}6\cdot 0{,}6 = 0{,}46. \blacktriangleright$$

▷ **Пример 1.29.** Экзаменационный билет для письменного экзамена состоит из 10 вопросов — по 2 вопроса из 20 по каждой из пяти тем, представленных в билете. По каждой теме студент подготовил лишь половину всех вопросов. Какова вероятность того, что студент сдаст экзамен, если для этого необходимо ответить хотя бы на один вопрос по каждой из пяти тем в билете?

Р е ш е н и е. Обозначим события:

A_1, A_2 — студент подготовил 1-й, 2-й вопросы билета по каждой теме;

B_i — студент подготовил хотя бы один вопрос билета из двух по i-й теме ($i = 1, 2, ..., 5$);

C — студент сдал экзамен.

В силу условия $C = B_1B_2B_3B_4B_5$. Полагая ответы студента по разным темам независимыми, по теореме умножения вероятностей (**1.24**)

$$P(C) = P(B_1)P(B_2)P(B_3)P(B_4)P(B_5).$$

Так как вероятности $P(B_i)$ ($i = 1, 2, ..., 5$) равны, то $P(C) = (P(B_i))^5$.

Вероятность $P(B_i)$ можно найти по формуле (**1.27**) (или (**1.25**)):

$$P(B_i) = P(A_1+A_2) = 1-P(A_1A_2) =$$

$$=1-P(A_1)P_{A_1}(A_2)=1-\frac{10}{20}\cdot\frac{9}{19}=0{,}763.$$

Теперь $P(C) = 0{,}763^5 = 0{,}259$. ▶

▷**Пример 1.30.** При включении зажигания двигатель начнет работать с вероятностью 0,6. Найти вероятность того, что: а) двигатель начнет работать при третьем включении зажигания; б) для запуска двигателя придется включать зажигание не более трех раз.

Р е ш е н и е. а) Обозначим события:

A — двигатель начнет работать при каждом включении зажигания;

B — то же при третьем включении зажигания.

Очевидно, что $B = \overline{A}\,\overline{A}A$ и $P(B) = P(\overline{A})P(\overline{A})P(A) = 0{,}4\cdot 0{,}4\cdot 0{,}6 = 0{,}096$.

б) Пусть событие C — для запуска двигателя придется включать зажигание не более трех раз. Очевидно, событие C наступит, если двигатель начнет работать при 1-м включении, или при 2-м, или при 3-м включении, т.е. $C = A + AA + A\,AA$. Следовательно,

$$P(C) = P(A) + P(A)P(A) + P(A)P(A)\cdot P(A) =$$
$$= 0{,}6 + 0{,}4\cdot 0{,}6 + 0{,}4\cdot 0{,}4\cdot 0{,}6 = 0{,}936. \blacktriangleright$$

▷**Пример 1.31.** Среди билетов денежно-вещевой лотереи половина выигрышных. Сколько лотерейных билетов нужно купить, чтобы с вероятностью, не меньшей 0,999, быть уверенным в выигрыше хотя бы по одному билету?

Р е ш е н и е. Пусть вероятность события A_i — выигрыша по i-му билету равна p, т.е. $P(A_i) = p$. Тогда вероятность выигрыша хотя бы по одному из n приобретенных билетов, т.е. вероятность суммы независимых событий $A_1, A_2, ..., A_i, ..., A_n$ определится по формуле (**1.29**):

$$P(A_1 + A_2 + ... + A_n) = 1 - (1 - p)^n.$$

По условию $1 - (1 - p)^n \geq \mathcal{P}$, где $\mathcal{P} = 0{,}999$, откуда

$$(1 - p)^n \leq 1 - \mathcal{P}.$$

Логарифмируя обе части неравенства, имеем

$$n\lg(1-p)\le \lg(1-\mathscr{P}).$$

Учитывая, что $\lg(1-p)$ — величина отрицательная, получим

$$n\ge\frac{\lg(1-\mathscr{P})}{\lg(1-p)}. \qquad \textbf{(1.30)}$$

По условию $p=0{,}5$, $\mathscr{P}=0{,}999$. По формуле **(1.30)**

$n\ge\frac{\lg 0{,}001}{\lg 0{,}5}=9{,}96$, т.е. $n\ge 10$ и необходимо купить не менее 10 лотерейных билетов.

(Задачу можно решить, не прибегая к логарифмированию, путем подбора целого числа n, при котором выполняется неравенство $(1-p)^n\le 1-\mathscr{P}$, т.е. в данном случае $\left(\frac{1}{2}\right)^n\le 0{,}001$; так, еще при $n=9$ $\left(\frac{1}{2}\right)^9=\frac{1}{512}>0{,}001$, а уже при $n=10$ $\left(\frac{1}{2}\right)^{10}=\frac{1}{1024}\le 0{,}001$, т.е. $n\ge 10$). ▶

▷**Пример 1.32.** Два игрока поочередно бросают игральную кость. Выигрывает тот, у которого первым выпадет «6 очков». Какова вероятность выигрыша для игрока, бросающего игральную кость первым? Вторым?

Р е ш е н и е. Обозначим события:

A_i — выпадение 6 очков при i-м бросании игральной кости ($i=1, 2, ...$);

B — выигрыш игры игроком, бросающим игральную кость первым.

Имеем $P(A_i)=1/6$, $P(\overline{A}_i)=5/6$ при любом i.

Событие B можно представить в виде суммы вариантов:

$B=A_1+\overline{A}_1\overline{A}_2A_3+\overline{A}_1\overline{A}_2\overline{A}_3\overline{A}_4A_5+...$ Поэтому

$$P(B)=P(A_1)+P(\overline{A}_1\overline{A}_2A_3)+P(\overline{A}_1\overline{A}_2\overline{A}_3\overline{A}_4A_5)+...=$$

$$=\frac{1}{6}+\left(\frac{5}{6}\right)^2\cdot\frac{1}{6}+\left(\frac{5}{6}\right)^4\cdot\frac{1}{6}+... .$$

По формуле суммы геометрического ряда с первым членом $a=1/6$ и знаменателем $q=(5/6)^2<1$

$$P(B)=\frac{a}{1-q}=\frac{1/6}{1-(5/6)^2}=\frac{6}{11}=0{,}545.$$

Вероятность $P(\overline{B})$ выигрыша игры игроком, бросающим игральную кость вторым, равна

$$P(\overline{B})=1-P(B)=1-6/11=5/11=0{,}455,$$

т.е. существенно меньше, чем игроком, бросающим игральную кость первым. ▶

▷ **Пример 1.33.** Вероятность попадания в мишень при каждом выстреле для 1-го стрелка равна 0,7, а для 2-го — 0,8. Оба они делают по одному выстрелу по мишени, а затем каждый из стрелков стреляет еще раз, если при первом сделанном им выстреле он промахнулся. Найти вероятность того, что в мишени ровно 2 пробоины.

Р е ш е н и е. Пусть события:

A_i, B_i — попадание в цель соответственно 1-м и 2-м стрелком при i-м выстреле ($i=1, 2$);

C — в мишени ровно 2 пробоины.

Событие C произойдет, если:

- у каждого стрелка по одному попаданию с первого раза;
- у 1-го стрелка — попадание (при одном выстреле), у 2-го стрелка промах и попадание;
- у 1-го стрелка — промах и попадание, у 2-го стрелка — попадание (при одном выстреле);
- у каждого стрелка — промах и попадание после двух выстрелов.

Итак,

$$C=A_1B_1+A_1\overline{B}_1B_2+\overline{A}_1B_1A_2+\overline{A}_1\overline{B}_1A_2B_2.$$

Используя теоремы сложения для несовместных и умножения для независимых событий, получим

$$P(C)=P(A_1B_1+A_1\overline{B}_1B_2+\overline{A}_1B_1A_2+\overline{A}_1\overline{B}_1A_2B_2)=$$

$$=0{,}7\cdot 0{,}8+0{,}7\cdot 0{,}2\cdot 0{,}8+0{,}3\cdot 0{,}8\cdot 0{,}7+0{,}3\cdot 0{,}2\cdot 0{,}7\cdot 0{,}8=0{,}8736.$$ ▶

1.11. Формула полной вероятности. Формула Байеса

Следствием двух основных теорем теории вероятностей — теоремы сложения и теоремы умножения — являются формула полной вероятности и формула Байеса.

Теорема. *Если событие F может произойти только при условии появления одного из событий (гипотез) $A_1, A_2,..., A_n$, образую-*

щих полную группу, то вероятность события F равна сумме произведений вероятностей каждого из этих событий (гипотез) на соответствующие условные вероятности события F:

$$P(F)=\sum_{i=1}^{n} P(A_i)P_{A_i}(F). \qquad (1.31)$$

□ По условию события (гипотезы) A_1, A_2, ..., A_n образуют полную группу, следовательно, они единственно возможные и несовместные.

Так как гипотезы A_1, A_2, ..., A_n — единственно возможные, а событие F по условию теоремы может произойти только вместе с одной из гипотез, то

$$F = A_1F + A_2F + ... + A_nF.$$

В силу того что гипотезы A_1, A_2, ..., A_n несовместны, можно применить теорему сложения вероятностей:

$$P(F) = P(A_1F) + P(A_2F) + ... + P(A_nF) = \sum_{i=1}^{n} P(A_iF).$$

По теореме умножения $P(A_iF) = P(A_i)\cdot P_{A_i}(F)$, откуда и получается утверждение **(1.31)**.■

Следствием теоремы умножения и формулы полной вероятности является **формула Байеса**.

Она применяется, когда событие F, которое может появиться только с одной из гипотез $A_1,A_2,...,A_n$, образующих полную группу событий, произошло и необходимо произвести количественную переоценку *априорных* вероятностей этих гипотез $P(A_1)$, $P(A_2),...,P(A_n)$, известных до испытания, т.е. надо найти *апостериорные* (получаемые после проведения испытания) условные вероятности гипотез $P_F(A_1),P_F(A_2),...,P_F(A_n)$.

□ Для получения искомой формулы запишем теорему умножения вероятностей событий F и A_i в двух формах:

$$P(FA_i) = P(F)P_F(A_i) = P(A_i)\cdot P_{A_i}(F),$$

откуда

$$P_F(A_i) = \frac{P(A_i)P_{A_i}(F)}{P(F)}, \qquad (1.32)$$

или с учетом **(1.31)**

$$P_F(A_i) = \frac{P(A_i)P_{A_i}(F)}{\sum_{i=1}^{n} P(A_i)P_{A_i}(F)}.■ \qquad (1.33)$$

Формула **(1.33)** называется *формулой Байеса*.

Значение формулы Байеса состоит в том, что при наступлении события F, т.е. по мере получения новой информации, мы можем проверять и корректировать выдвинутые до испытания гипотезы. Такой подход, называемый *байесовским*, дает возможность корректировать управленческие решения в экономике, оценки неизвестных параметров распределения изучаемых признаков в статистическом анализе и т.п.

▷**Пример 1.34.** В торговую фирму поступили телевизоры от трех поставщиков в отношении 1:4:5. Практика показала, что телевизоры, поступающие от 1-го, 2-го и 3-го поставщиков, не потребуют ремонта в течение гарантийного срока соответственно в 98, 88 и 92% случаев.

1) Найти вероятность того, что поступивший в торговую фирму телевизор не потребует ремонта в течение гарантийного срока. 2) Проданный телевизор потребовал ремонта в течение гарантийного срока. От какого поставщика вероятнее всего поступил этот телевизор?

Р е ш е н и е. 1) Обозначим события:

A_i — телевизор поступил в торговую фирму от i-го поставщика ($i = 1, 2, 3$);

F — телевизор не потребует ремонта в течение гарантийного срока.

По условию

$$P(A_1)=\frac{1}{1+4+5}=0,1; \qquad P_{A_1}(F)=0,98;$$

$$P(A_2)=\frac{4}{1+4+5}=0,4; \qquad P_{A_2}(F)=0,88;$$

$$P(A_3)=\frac{5}{1+4+5}=0,5; \qquad P_{A_3}(F)=0,92.$$

По формуле полной вероятности **(1.31)**

$$P(F)=0,1\cdot 0,98+0,4\cdot 0,88+0,5\cdot 0,92=0,91.$$

2) Событие $\overline{F}$ — телевизор потребует ремонта в течение гарантийного срока; $P(\overline{F}) = 1 - P(F) = 1 - 0,91 = 0,09$.

По условию

$$P_{A_1}(\overline{F})=1-0,98=0,02,$$

$$P_{A_2}(\overline{F})=1-0,88=0,12,$$

$$P_{A_3}(\overline{F}) = 1 - 0{,}92 = 0{,}08.$$

По формуле Байеса **(1.32)**

$$P_{\overline{F}}(A_1) = \frac{0{,}1 \cdot 0{,}02}{0{,}09} = 0{,}022; \quad P_{\overline{F}}(A_2) = \frac{0{,}4 \cdot 0{,}12}{0{,}09} = 0{,}533;$$

$$P_{\overline{F}}(A_3) = \frac{0{,}5 \cdot 0{,}08}{0{,}09} = 0{,}444.$$

Таким образом, после наступления события $\overline{F}$ вероятность гипотезы A_2 увеличилась с $P(A_2) = 0{,}4$ до максимальной $P_{\overline{F}}(A_2) = 0{,}533$, а гипотезы A_3 — уменьшилась от максимальной $P(A_3) = 0{,}5$ до $P_{\overline{F}}(A_3) = 0{,}444$; если ранее (до наступления события F) наиболее вероятной была гипотеза A_3, то теперь, в свете новой информации (наступления события F), наиболее вероятна гипотеза A_2 — поступление данного телевизора от 2-го поставщика. ▶

▷ **Пример 1.35.** Известно, что в среднем 95% выпускаемой продукции удовлетворяют стандарту. Упрощенная схема контроля признает пригодной продукцию с вероятностью 0,98, если она стандартна, и с вероятностью 0,06, если она нестандартна. Определить вероятность того, что: 1) взятое наудачу изделие пройдет упрощенный контроль; 2) изделие стандартное, если оно: а) прошло упрощенный контроль; б) дважды прошло упрощенный контроль.

Р е ш е н и е. 1) Обозначим события:

A_1, A_2 — взятое наудачу изделие соответственно стандартное или нестандартное;

F — изделие прошло упрощенный контроль.

По условию

$$P(A_1) = 0{,}95, \quad P(A_2) = 0{,}05, \quad P_{A_1}(F) = 0{,}98; \quad P_{A_2}(F) = 0{,}06.$$

Вероятность того, что взятое наудачу изделие пройдет упрощенный контроль, по формуле полной вероятности **(1.31)**:

$$P(F) = 0{,}95 \cdot 0{,}98 + 0{,}05 \cdot 0{,}06 = 0{,}934.$$

2. а) Вероятность того, что изделие, прошедшее упрощенный контроль, стандартное, по формуле Байеса **(1.32)**:

$$P_F(A_1) = \frac{0{,}95 \cdot 0{,}98}{0{,}934} = 0{,}997.$$

б) Пусть событие F^* — изделие дважды прошло упрощенный контроль. Тогда по теореме умножения вероятностей

$$P_{A_1}(F^*) = 0{,}98 \cdot 0{,}98 = 0{,}9604 \quad \text{и} \quad P_{A_2}(F^*) = 0{,}06 \cdot 0{,}06 = 0{,}0036.$$

По формуле Байеса **(1.33)**

$$P_{F^*}(A_1) = \frac{0{,}95 \cdot 0{,}9604}{0{,}95 \cdot 0{,}9604 + 0{,}05 \cdot 0{,}0036} = 0{,}9998.$$

Так как

$$P_{F^*}(A_2) = 1 - P_{F^*}(A_1) = 1 - 0{,}9998 = 0{,}0002$$

очень мала, то гипотезу A_2 о том, что изделие, дважды прошедшее упрощенный контроль, нестандартное, следует отбросить как практически невозможное событие. ▶

▷ **Пример 1.36.** Два стрелка независимо друг от друга стреляют по мишени, делая каждый по одному выстрелу. Вероятность попадания в мишень для первого стрелка равна 0,8; для второго — 0,4. После стрельбы в мишени обнаружена одна пробоина. Какова вероятность того, что она принадлежит: а) 1-му стрелку; б) 2-му стрелку?

Р е ш е н и е. Обозначим события:

A_1 — оба стрелка не попали в мишень;

A_2 — оба стрелка попали в мишень;

A_3 — 1-й стрелок попал в мишень, 2-й нет;

A_4 — 1-й стрелок не попал в мишень, 2-й попал;

F — в мишени одна пробоина (одно попадание).

Найдем вероятности гипотез и условные вероятности события F для этих гипотез:

$$P(A_1) = 0{,}2 \cdot 0{,}6 = 0{,}12, \quad P_{A_1}(F) = 0;$$

$$P(A_2) = 0{,}8 \cdot 0{,}4 = 0{,}32, \quad P_{A_2}(F) = 0;$$

$$P(A_3) = 0{,}8 \cdot 0{,}6 = 0{,}48, \quad P_{A_3}(F) = 1;$$

$$P(A_4) = 0{,}2 \cdot 0{,}4 = 0{,}08, \quad P_{A_4}(F) = 1.$$

Теперь по формуле Байеса **(1.33)**

$$P_F(A_3) = \frac{0{,}48 \cdot 1}{0{,}12 \cdot 0 + 0{,}32 \cdot 0 + 0{,}48 \cdot 1 + 0{,}08 \cdot 1} = \frac{6}{7} = 0{,}857,$$

$$P_F(A_4)=\frac{0{,}08\cdot 1}{0{,}12\cdot 0+0{,}32\cdot 0+0{,}48\cdot 1+0{,}08\cdot 1}=\frac{1}{7}=0{,}143,$$

т.е. вероятность того, что попал в цель 1-й стрелок *при наличии одной пробоины*, в 6 раз выше, чем для второго стрелка. ▶

1.12. Теоретико-множественная трактовка основных понятий и аксиоматическое построение теории вероятностей

Приведем **теоретико-множественную трактовку** основных понятий теории вероятностей, рассмотренных выше.

Пусть Ω — множество всех возможных исходов некоторого испытания (опыта, эксперимента). Каждый элемент ω множества Ω, т.е. $\omega\in\Omega$, называют *элементарным событием* или *элементарным исходом*, а само множество Ω — *пространством элементарных событий*. Любое событие A рассматривается как некоторое подмножество (часть) множества Ω, т.е. $A\subset\Omega$.

Так, в примере **1.1** при бросании игральной кости возможны 6 элементарных исходов (событий): ω_1 — выпадение 1 очка, ω_2 — выпадение 2 очков, ... , ω_6 — выпадение 6 очков, т.е. пространство элементарных событий $\Omega=\{\omega_1,\omega_2,\omega_3,\omega_4,\omega_5,\omega_6\}$. Событие A, состоящее в выпадении четного числа очков, есть $A=\{\omega_2,\omega_4,\omega_6\}$.

В задаче о встрече, приведенной в примере **1.2**, возможно бесконечное несчетное множество элементарных исходов (событий) — точек (x, y) квадрата $OKLM$, координаты x и y которых равны моментам прихода к месту встречи двух лиц (рис. 1.2), т.е. пространство элементарных событий Ω — квадрат $OKLM$. Событие A, состоящее в том, что встреча двух лиц произойдет, есть заштрихованная область g на рисунке — часть квадрата, т.е. подмножество пространства Ω: $A\subset\Omega$.

Само пространство элементарных событий Ω представляет собой событие, происходящее всегда (при любом элементарном исходе ω), и называется *достоверным* событием. Таким образом, Ω выступает в двух качествах: множества всех элементарных исходов и достоверного события.

Ко всему пространству Ω элементарных событий добавляется еще пустое множество $\varnothing$, рассматриваемое как событие и называемое *невозможным* событием.

Суммой нескольких событий A_1, A_2, ..., A_n называется объединение множеств $A_1\cup A_2\cup\ldots\cup A_n$.

Произведением нескольких событий A_1, A_2, ..., A_n называется пересечение множеств $A_1\cap A_2\cap\ldots\cap A_n$.

Событием $\overline{A}$, противоположным событию A, называется дополнение множества A до Ω, т.е. $\Omega\backslash A$.

На диаграммах Венна (см. § 1.7, рис. 1.3 *в*, *г*, *д*, *е*) представлены сумма $A+B$, произведение AB двух событий и события $\overline{A}$, $\overline{B}$, противоположные событиям A, B.

Несколько событий A_1, A_2, ..., A_n образуют *полную группу* (*полную систему*), если их сумма представляет все пространство элементарных событий, а сами события несовместные, т.е. $\sum_{i=1}^{n}A_i=\Omega$ и $A_iA_j=\varnothing(i\neq j)$.

Таким образом, *под операциями над событиями понимаются операции над соответствующими множествами.* В табл. 1.1 показано соответствие терминов теории множеств и теории вероятностей.

Таблица 1.1

Обозначения	*Термины*	
	Теории множеств	*Теории вероятностей*
Ω	Множество, пространство	Пространство элементарных событий, достоверное событие
ω	Элемент множества	Элементарное событие (элементарный исход)
A, B	Подмножество A, B	Событие A, B
$A+B=A\cup B$	Объединение (сумма) множеств A и B	Сумма событий A и B
$AB=A\cap B$	Пересечение множеств A и B	Произведение событий A и B
$\varnothing$	Пустое множество	Невозможное событие
$\overline{A}$	Дополнение множества A	Противоположное для A событие
$AB=A\cap B=\varnothing$	Множества A и B не пересекаются	События A и B несовместны
$A=B$	Множества A и B равны	События A и B равносильны
$A\subset B$	A есть подмножество B	Событие A влечет за собой событие B

На основе изложенной трактовки событий как множеств перейдем к **аксиоматическому построению** теории вероятностей.

Необходимость формально логического обоснования теории вероятностей, ее аксиоматического построения возникла в связи с развитием самой теории вероятностей как математической науки и ее приложений в различных областях.

Такие сформировавшиеся науки, как геометрия, теоретическая механика, теория множеств, строятся аксиоматически. Фундаментом каждой служит ряд аксиом, являющихся обобщением многовекового человеческого опыта, а само здание науки строится на основе строгих логических рассуждений без обращения к наглядным представлениям.

Аксиоматика теории вероятностей исходит от основных свойств вероятности событий, к которым применимо классическое или статистическое определение вероятности. Аксиоматическое определение вероятности как частные случаи включает в себя и классическое, и статистическое определения и преодолевает недостатки каждого из них.

Впервые идея аксиоматического построения вероятностей была высказана российским академиком С.Н. Бернштейном, исходившим из качественного сравнения событий по их большей или меньшей вероятности. В начале 1930-х гг. академик А.Н. Колмогоров разработал иной подход, связывающий теорию вероятностей с современной метрической теорией функций и теорией множеств, который в настоящее время является общепринятым.

Сформулируем а к с и о м ы теории вероятностей. Каждому событию A поставим в соответствие некоторое число, называемое *вероятностью события A*, т.е. $P(A)$. Так как любое событие есть *множество*, то вероятность события есть *функция множества*.

Вероятность события должна удовлетворять следующим а к с и о м а м:

Р.1. *Вероятность любого события неотрицательна*:

$$P(A) \geq 0.$$

Р.2. *Вероятность достоверного события равна* 1:

$$P(\Omega) = 1.$$

Р.3. *Вероятность суммы несовместных событий равна сумме вероятностей этих событий, т.е. если* $A_iA_j = \varnothing$ $(i \neq j)$, *то*

$$P(A_1+A_2+...+A_n) = P(A_1)+P(A_2)+...+P(A_n).$$

Для классического определения вероятности свойства, выраженные аксиомами **Р.2**, **Р.3**, не нужно постулировать, так как эти свойства были нами доказаны выше.

Из аксиом **Р.1**, **Р.2**, **Р.3** можно вывести основные с в о й с т в а вероятностей, известные нам из предыдущего изложения:

1. $P(A) = 1 - P(A)$.
2. $P(\varnothing)=0$.
3. $0 \leq P(A) \leq 1$.
4. $P(A) \leq P(B)$, если $A \subset B$.
5. $P(A+B)=P(A)+P(B)-P(AB)$.
6. $P(A+B) \leq P(A)+P(B)$.

В случае произвольного (не обязательно конечного) пространства элементарных событий Ω аксиому **Р.3** необходимо заменить более сильной, расширенной аксиомой сложения **Р.3'** (которую нельзя вывести из аксиомы **Р.3**).

Если имеется счетное[1] множество несовместных событий $A_1, A_2, ..., A_n, ...$, $(A_iA_j = \varnothing$ при $i \neq j)$, то

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

Аксиомы теории вероятностей позволяют вычислить вероятности любых событий (подмножеств пространства Ω) через вероятности элементарных событий (если их конечное или счетное число). Вопрос о том, откуда берутся вероятности элементарных событий, при аксиоматическом построении теории вероятностей не рассматривается. На практике они определяются с помощью классического определения (если испытание сводится к схеме случаев) или статистического определения.

Сформулированные аксиомы не определяют условной вероятности одного события относительно другого, которая вводится по определению.

О п р е д е л е н и е. ***Условная вероятность события*** *B относительно события A есть отношение вероятности произведения этих событий к вероятности события A*, т.е.

$$P_A(B) = \frac{P(AB)}{P(A)}, \qquad \textbf{(1.34)}$$

если $P(A) \neq 0$.

[1] Напомним, что множество называется *счетным*, если его элементы можно перенумеровать натуральными числами.

Из этого определения автоматически следует теорема (правило) умножения вероятностей для любых событий.

В последующих главах мы по-прежнему будем ссылаться на теоремы (правила) сложения и умножения вероятностей, хотя правильнее было бы говорить об аксиоме сложения, определении условной вероятности.

В более полных курсах теории вероятностей рассматривается понятие *вероятностного пространства*, определяемого *тройкой компонент (символов)* (Ω, S, P), где Ω — пространство элементарных событий, S — σ(сигма)-алгебра событий, P — вероятность. Первую (Ω) и третью (P) компоненты вероятностного пространства мы уже рассмотрели в данном параграфе.

Вторая компонента (S) вероятностного пространства — σ-алгебра событий — представляет собой некоторую систему подмножеств пространства элементарных исходов (событий) Ω. Если Ω конечно или счетно, то любое подмножество элементарных исходов является событием, а σ-алгебра есть система всех этих подмножеств. Если же Ω более чем счетно, то оказывается, что не каждое произвольное подмножество Ω может быть названо событием. Причина этого заключается в существовании так называемых *неизмеримых* подмножеств. Поэтому в этом случае под событием понимается уже не любое подмножество пространства Ω, а только подмножество из выделенного класса S, а σ-алгебра есть система таких подмножеств. Рассмотрение указанных вопросов выходит за рамки данной книги.

Упражнения

1.37. Слово составлено из карточек, на каждой из которых написана одна буква. Карточки смешивают и вынимают без возврата по одной. Найти вероятность того, что карточки с буквами вынимаются в порядке следования букв заданного слова: а) «событие»; б) «статистика».

1.38. Пятитомное собрание сочинений расположено на полке в случайном порядке. Какова вероятность того, что книги стоят слева направо в порядке нумерации томов (от 1 до 5)?

1.39. Среди 25 студентов, из которых 15 девушек, разыгрываются четыре билета, причем каждый может выиграть только один билет. Какова вероятность того, что среди обладателей билета окажутся: а) четыре девушки; б) четыре юноши; в) три юноши и одна девушка?

1.40. Из 20 сбербанков 10 расположены за чертой города. Для обследования случайным образом отобрано 5 сбербанков. Какова вероятность того, что среди отобранных окажется в черте города: а) 3 сбербанка; б) хотя бы один?

1.41. Из ящика, содержащего 5 пар обуви, из которых три пары мужской, а две пары женской обуви, перекладывают наудачу 2 пары обуви в другой ящик, содержащий одинаковое количество пар женской и мужской обуви. Какова вероятность того, что во втором ящике после этого окажется одинаковое количество пар мужской и женской обуви?

1.42. В магазине имеются 30 телевизоров, причем 20 из них импортных. Найти вероятность того, что среди 5 проданных в течение дня телевизоров окажется не менее 3 импортных телевизоров, предполагая, что вероятности покупки телевизоров разных марок одинаковы.

1.43. Наудачу взятый телефонный номер состоит из 5 цифр. Какова вероятность того, что в нем все цифры: а) различные; б) одинаковые; в) нечетные? Известно, что номер телефона не начинается с цифры ноль.

1.44. Для проведения соревнования 16 волейбольных команд разбиты по жребию на две подгруппы (по восемь команд в каждой). Найти вероятность того, что две наиболее сильные команды окажутся: а) в разных подгруппах; б) в одной подгруппе.

1.45. Студент знает 20 из 25 вопросов программы. Зачет считается сданным, если студент ответит не менее чем на 3 из 4 поставленных в билете вопросов. Взглянув на первый вопрос билета, студент обнаружил, что он его знает. Какова вероятность того, что студент: а) сдаст зачет; б) не сдаст зачет?

1.46. У сборщика имеются 10 деталей, мало отличающихся друг от друга, из них четыре — первого, по две — второго, третьего и четвертого видов. Какова вероятность того, что среди шести взятых одновременно деталей три окажутся первого вида, два — второго и одна — третьего?

1.47. Найти вероятность того, что из 10 книг, расположенных в случайном порядке, 3 определенные книги окажутся рядом.

1.48. В старинной игре в кости необходимо было для выигрыша получить при бросании трех игральных костей сумму очков, превосходящую 10. Найти вероятности: а) выпадения 11 очков; б) выигрыша.

1.49. На фирме работают 8 аудиторов, из которых 3 — высокой квалификации, и 5 программистов, из которых 2 — высокой квалификации. В командировку надо отправить группу из 3 аудиторов и 2 программистов. Какова вероятность того, что в этой группе окажется по крайней мере 1 аудитор высокой квалификации и хотя бы 1 программист высокой квалификации, если каждый специалист имеет равные возможности поехать в командировку?

1.50. Два лица условились встретиться в определенном месте между 18 и 19 ч и договорились, что пришедший первым ждет другого в течение 15 мин., после чего уходит. Найти вероятность их встречи, если приход каждого в течение указанного часа может произойти в любое время и моменты прихода независимы.

1.51. Какова вероятность того, что наудачу брошенная в круг точка окажется внутри вписанного в него квадрата?

1.52. При приеме партии изделий подвергается проверке половина изделий. Условие приемки — наличие брака в выборке менее 2%. Вычислить вероятность того, что партия из 100 изделий, содержащая 5% брака, будет принята.

1.53. По результатам проверки контрольных работ оказалось, что в первой группе получили положительную оценку 20 студентов из 30, а во второй — 15 из 25. Найти вероятность того, что наудачу выбранная работа, имеющая положительную оценку, написана студентом первой группы.

1.54. Экспедиция издательства отправила газеты в три почтовых отделения. Вероятность своевременной доставки газет в первое отделение равна 0,95, во второе отделение — 0,9 и в третье — 0,8. Найти вероятность следующих событий: а) только одно отделение получит газеты вовремя; б) хотя бы одно отделение получит газеты с опозданием.

1.55. Прибор, работающий в течение времени t, состоит из трех узлов, каждый из которых независимо от других может за это время выйти из строя. Неисправность хотя бы одного узла выводит прибор из строя целиком. Вероятность безотказной работы в течение времени t первого узла равна 0,9, второго — 0,95, третьего — 0,8. Найти вероятность того, что в течение времени t прибор выйдет из строя.

1.56. Студент разыскивает нужную ему формулу в трех справочниках. Вероятность того, что формула содержится в первом, втором и третьем справочниках, равна соответственно 0,6, 0,7 и 0,8. Найти вероятность того, что эта формула содержится не менее, чем в двух справочниках.

1.57. Произведено три выстрела по цели из орудия. Вероятность попадания при первом выстреле равна 0,75; при втором — 0,8; при третьем — 0,9. Определить вероятность того, что будет: а) три попадания; б) хотя бы одно попадание.

1.58. Вероятность своевременного выполнения студентом контрольной работы по каждой из трех дисциплин равна соответственно 0,6, 0,5 и 0,8. Найти вероятность своевременного выполнения контрольной работы студентом: а) по двум дисциплинам; б) хотя бы по двум дисциплинам.

1.59. Мастер обслуживает 4 станка, работающих независимо друг от друга. Вероятность того, что первый станок в течение смены потребует внимания рабочего, равна 0,3, второй — 0,6, третий — 0,4 и четвертый — 0,25. Найти вероятность того, что в течение смены хотя бы один станок не потребует внимания мастера.

1.60. Контролер ОТК, проверив качество сшитых 20 пальто, установил, что 16 из них первого сорта, а остальные — второго. Найти вероятность того, что среди взятых наудачу из этой партии трех пальто одно будет второго сорта.

1.61. Среди 20 поступающих в ремонт часов 8 нуждаются в общей чистке механизма. Какова вероятность того, что среди взятых одновременно наудачу 3 часов по крайней мере двое нуждаются в общей чистке механизма?

1.62. Среди 15 лампочек 4 стандартные. Одновременно берут наудачу 2 лампочки. Найти вероятность того, что хотя бы одна из них нестандартная.

1.63. В коробке смешаны электролампы одинакового размера и формы: по 100 Вт — 7 штук, по 75 Вт — 13 штук. Вынуты наудачу 3 лампы. Какова вероятность того,

что: а) они одинаковой мощности; б) хотя бы две из них по 100 Вт?

1.64. В коробке 10 красных, 3 синих и 7 желтых карандашей. Наудачу вынимают 3 карандаша. Какова вероятность того, что они все: а) разных цветов; б) одного цвета?

1.65. Брак в продукции завода вследствие дефекта A составляет 4%, а вследствие дефекта B — 3,5%. Годная продукция завода составляет 95%. Найти вероятность того, что: а) среди продукции, не обладающей дефектом A, встретится дефект B; б) среди забракованной по признаку A продукции встретится дефект B.

1.66. Пакеты акций, имеющихся на рынке ценных бумаг, могут дать доход владельцу с вероятностью 0,5 (для каждого пакета). Сколько пакетов акций различных фирм нужно приобрести, чтобы с вероятностью, не меньшей 0,96875, можно было ожидать доход хотя бы по одному пакету акций?

1.67. Сколько раз нужно провести испытание, чтобы с вероятностью, не меньшей P, можно было утверждать, что по крайней мере один раз произойдет событие, вероятность которого в каждом испытании равна p? Дать ответ при $p = 0{,}4$ и $P = 0{,}8704$.

1.68. На полке стоят 10 книг, среди которых 3 книги по теории вероятностей. Наудачу берутся три книги. Какова вероятность, что среди отобранных хотя бы одна книга по теории вероятностей?

1.69. На связке 5 ключей. К замку подходит только один ключ. Найти вероятность того, что потребуется не более двух попыток открыть замок, если опробованный ключ в дальнейших испытаниях не участвует.

1.70. В магазине продаются 10 телевизоров, 3 из них имеют дефекты. Какова вероятность того, что посетитель купит телевизор, если для выбора телевизора без дефектов понадобится не более трех попыток?

1.71. Радист трижды вызывает корреспондента. Вероятность того, что будет принят первый вызов, равна 0,2, второй — 0,3, третий — 0,4. События, состоящие в том, что данный вызов будет услышан, независимы. Найти вероятность того, что корреспондент услышит вызов радиста.

1.72. Страховая компания разделяет застрахованных по классам риска: I класс — малый риск, II класс — средний, III класс — большой риск. Среди этих клиентов 50% — первого класса риска, 30% — второго и 20% — третьего. Вероятность необходимости выплачивать страховое вознаграждение для первого класса риска равна 0,01, второго — 0,03, третьего — 0,08. Какова вероятность того, что: а) застрахованный получит денежное вознаграждение за период страхования; б) получивший денежное вознаграждение застрахованный относится к группе малого риска?

1.73. В данный район изделия поставляются тремя фирмами в соотношении 5:8:7. Среди продукции первой фирмы стандартные изделия составляют 90%, второй — 85%, третьей — 75%. Найти вероятность того, что: а) приобретенное изделие окажется нестандартным; б) приобретенное изделие оказалось стандартным. Какова вероятность того, что оно изготовлено третьей фирмой?

1.74. Два стрелка сделали по одному выстрелу в мишень. Вероятность попадания в мишень для первого стрелка равна 0,6, а для второго — 0,3. В мишени оказалась одна пробоина. Найти вероятность того, что она принадлежит первому стрелку.

1.75. Вся продукция цеха проверяется двумя контролерами, причем первый контролер проверяет 55% изделий, а второй — остальные. Вероятность того, что первый контролер пропустит нестандартное изделие, равна 0,01, второй — 0,02. Взятое наудачу изделие, маркированное как стандартное, оказалось нестандартным. Найти вероятность того, что это изделие проверялось вторым контролером.

1.76. Вероятность изготовления изделия с браком на данном предприятии равна 0,04. Перед выпуском изделие подвергается упрощенной проверке, которая в случае бездефектного изделия пропускает его с вероятностью 0,96, а в случае изделия с дефектом — с вероятностью 0,05. Определить: а) какая часть изготовленных изделий выходит с предприятия; б) какова вероятность того, что изделие, выдержавшее упрощенную проверку, бракованное?

1.77. В одной урне 5 белых и 6 черных шаров, а в другой — 4 белых и 8 черных шаров. Из первой урны случайным образом вынимают 3 шара и опускают во вторую урну. После этого из второй урны также случайно вынимают 4 шара. Найти вероятность того, что все шары, вынутые из второй урны, белые.

1.78. Из n экзаменационных билетов студент A подготовил только m ($m<n$). В каком случае вероятность вытащить на экзамене «хороший» для него билет выше: когда он берет наудачу билет первым, или вторым,..., или k-м ($k<n$) по счету среди сдающих экзамен?

1.79. В лифт семиэтажного дома на первом этаже вошли три человека. Каждый из них с одинаковой вероятностью выходит на любом из этажей, начиная со второго. Найти вероятность того, что все пассажиры выйдут: а) на четвертом этаже; б) на одном и том же этаже; в) на разных этажах.

1.80. Батарея, состоящая из 3 орудий, ведет огонь по группе, состоящей из 5 самолетов. Каждое орудие выбирает себе цель случайно и независимо от других. Найти вероятность того, что все орудия будут стрелять: а) по одной и той же цели; б) по разным целям.

1.81. 20 человек случайным порядком рассаживаются за столом. Найти вероятность того, что два фиксированных лица A и B окажутся рядом, если: а) стол круглый; б) стол прямоугольный, а 20 человек рассаживаются случайно вдоль одной из его сторон.

1.82. Имеется коробка с девятью новыми теннисными мячами. Для игры берут три мяча; после игры их кладут обратно. При выборе мячей игранные от неигранных не отличаются. Какова вероятность того, что после трех игр в коробке не останется неигранных мячей?

1.83. Завод выпускает определенного типа изделия; каждое изделие имеет дефект с вероятностью 0,7. После изготовления изделие осматривается последовательно тремя контролерами, каждый из которых обнаруживает дефект с вероятностями 0,8; 0,85; 0,9 соответственно. В случае обнаружения дефекта изделие бракуется. Определить вероятность того, что изделие: 1) будет забраковано; 2) будет забраковано: а) вторым контролером; б) всеми контролерами.

1.84. Из полной колоды карт (52 карты) выбирают шесть карт; одну из них смотрят; она оказывается тузом, после чего ее смешивают с остальными выбранными картами. Найти вероятность того, что при втором извлечении карты из этих шести мы снова получим туз.

1.85. В урне два белых и три черных шара. Два игрока поочередно вынимают из урны по шару, не вкладывая их обратно. Выигрывает тот, кто раньше получит белый шар. Найти вероятность того, что выиграет первый игрок.

1.86. Производятся испытания прибора. При каждом испытании прибор выходит из строя с вероятностью 0,8. После первого выхода из строя прибор ремонтируется; после второго признается негодным. Найти вероятность того, что прибор окончательно выйдет из строя в точности при четвертом испытании.

1.87. Имеется 50 экзаменационных билетов, каждый из которых содержит два вопроса. Экзаменующийся знает ответ не на все 100 вопросов, а только на 60. Определить вероятность того, что экзамен будет сдан, если для этого достаточно ответить на оба вопроса из своего билета, или на один вопрос из своего билета, или на один (по выбору преподавателя) вопрос из дополнительного билета.

1.88. Прибор состоит из двух узлов: работа каждого узла безусловно необходима для работы прибора в целом. Надежность (вероятность безотказной работы в течение времени t) первого узла равна 0,8, второго — 0,9. Прибор испытывался в течение времени t, в результате чего обнаружено, что он вышел из строя (отказал). Найти вероятность того, что отказал только первый узел, а второй исправен.

1.89. В группе из 10 студентов, пришедших на экзамен, 3 подготовлено отлично, 4 — хорошо, 2 — посредственно и 1 — плохо. В экзаменационных билетах имеется 20 вопросов. Отлично подготовленный студент может ответить на все 20 вопросов, хорошо подготовленный — на 16, посредственно — на 10, плохо — на 5. Вызванный наугад студент ответил на три произвольно заданных вопроса. Найти вероятность того, что студент подготовлен: а) отлично; б) плохо.

Глава 2 Повторные независимые испытания

На практике часто приходится сталкиваться с задачами, которые можно представить в виде многократно повторяющихся испытаний при данном комплексе условий, в которых представляет интерес вероятность числа m наступлений некоторого события A в n испытаниях. Например, необходимо определить вероятность определенного числа попаданий в мишень при нескольких выстрелах, вероятность некоторого числа бракованных изделий в данной партии и т.д.

Если вероятность наступления события A в каждом испытании не меняется в зависимости от исходов других, то такие испытания называются *независимыми относительно события A.* Если независимые повторные испытания проводятся при одном и том же комплексе условий, то *вероятность наступления события A в каждом испытании одна и та же.* Описанная последовательность независимых испытаний получила название *схемы Бернулли.*

2.1. Формула Бернулли

Теорема. *Если вероятность p наступления события A в каждом испытании постоянна, то вероятность $P_{m,n}$ того, что событие A наступит m раз в n независимых испытаниях, равна*

$$P_{m,n} = C_n^m p^m q^{n-m}, \qquad (2.1)$$

где $q = 1 - p$.

□ Пусть A_i и $\overline{A}_i$ — соответственно появление и непоявление события A в i-м испытании ($i = 1, 2, ..., n$), а B_m — событие, состоящее в том, что в n независимых испытаниях событие A появилось m раз.

Представим событие B_m через элементарные события A_i.

Например, при $n = 3$, $m = 2$ событие

$$B_2 = A_1A_2\overline{A}_3 + A_1\overline{A}_2A_3 + \overline{A}_1A_2A_3,$$

т.е. событие A произойдет два раза в трех испытаниях, если оно произойдет в 1-м и 2-м испытаниях (и не произойдет в 3-м), или в 1-м и 3-м (и не произойдет во 2-м), или произойдет во 2-м и 3-м (и не произойдет в 1-м).

В общем виде

$$B_m = \underline{A_1}\underline{A_2}...\underline{A_m}\overline{A}_{m+1}...\overline{A}_n + A_1\overline{A}_2A_3...\overline{A}_{n-1}A_n + ... + A_1A_2...A_{n-m}A_{n-m+1}...A_n, \qquad (2.2)$$

т.е. каждый вариант появления события B_m (каждый член суммы (2.2)) состоит из m появлений события A и $n-m$ непоявлений, т.е. из m событий A и из $n-m$ событий $\overline{A}$ с различными индексами.

Число всех комбинаций (слагаемых суммы (2.2)) равно числу способов выбора из n испытаний m, в которых событие A произошло, т.е. числу сочетаний C_n^m. Вероятность каждой такой комбинации (каждого варианта появления события B_m) по теореме умножения для независимых событий равна p^mq^{n-m}, так как $p(A_i) = p$, $p(\overline{A}_i) = q$, $i = 1, 2, ..., n$. В связи с тем, что комбинации между собой несовместны, по теореме сложения вероятностей получим

$$P_{m,n} = P(B_m) = \underbrace{p^mq^{n-m} + ... + p^mq^{n-m}}_{C_n^m \text{ раз}} = C_n^m p^m q^{n-m}. \blacksquare$$

▷ **Пример 2.1.** Вероятность изготовления на автоматическом станке стандартной детали равна 0,8. Найти вероятности возможного числа появления бракованных деталей среди 5 отобранных.

Р е ш е н и е. Вероятность изготовления бракованной детали $p = 1 - 0,8 = 0,2$. Искомые вероятности находим по формуле Бернулли (2.1):

$P_{0,5} = C_5^0 \cdot 0,2^0 \cdot 0,8^5 = 0,32768$; $P_{1,5} = C_5^1 \cdot 0,2^1 \cdot 0,8^4 = 0,4096$;

$P_{2,5} = C_5^2 \cdot 0,2^2 \cdot 0,8^3 = 0,2048$; $P_{3,5} = C_5^3 \cdot 0,2^3 \cdot 0,8^2 = 0,0512$;

$P_{4,5} = C_5^4 \cdot 0,2^4 \cdot 0,8^1 = 0,0064$; $P_{5,5} = C_5^5 \cdot 0,2^5 \cdot 0,8^0 = 0,00032$.

Полученные вероятности изобразим графически точками с координатами $(m, P_{m,n})$. Соединяя эти точки, получим *многоугольник*, или *полигон*, *распределения вероятностей* (рис. 2.1). ▶

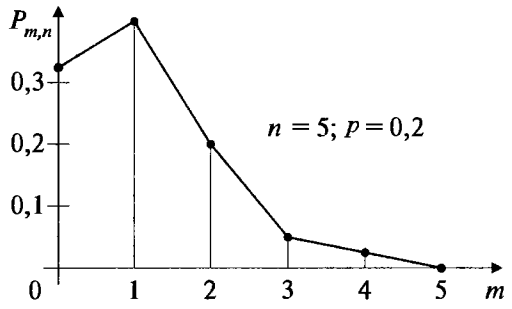


Рис. 2.1

Рассматривая многоугольник распределения вероятностей (рис. 2.1), мы видим, что есть такие значения m (в данном случае, одно — m_0=1), обладающие наибольшей вероятностью $P_{m,n}$.

Число m_0 наступления события A в n независимых испытаниях называется *наивероятнейшим*, если вероятность осуществления этого события $P_{m_0,n}$ по крайней мере не меньше вероятностей других событий $P_{m,n}$ при любом m.

Для нахождения m_0 запишем систему неравенств:

$$\begin{cases} P_{m_0,n} \geq P_{m_0+1,n}, \\ P_{m_0,n} \geq P_{m_0-1,n}. \end{cases} \qquad \textbf{(2.3)}$$

Решим первое неравенство системы **(2.3)**. Используя формулы Бернулли и числа сочетаний, запишем:

$$\frac{n!}{m_0!(n-m_0)!}p^{m_0}q^{n-m_0} \geq \frac{n!}{(m_0+1)!(n-m_0-1)!}p^{m_0+1}q^{n-m_0-1}.$$

Так как $(m_0+1)! = m_0!(m_0+1)$, $(n-m_0)! = (n-m_0-1)!(n-m_0)$, то получим после упрощений неравенство $\frac{1}{n-m_0}q \geq \frac{1}{m_0+1}p$, откуда $(m_0+1)\,q \geq (n-m_0)\,p$.

Теперь $m_0(p+q) \geq np - q$ или $m_0 \geq np - q$ (ибо $p+q=1$).

Решая второе неравенство системы **(2.3)**, получим аналогично: $m_0 \leq np + p$. Объединяя полученные решения двух неравенств, придем к двойному неравенству:

$$np - q \leq m_0 \leq np + p. \qquad \textbf{(2.4)}$$

Отметим, что так как разность $np + p - (np - q) = p + q = 1$, то всегда существует целое число m_0, удовлетворяющее неравенству **(2.4)**. При этом, если $np + p$ — целое число, то наивероятнейших чисел два: $m_0 = np + p$ и $m_0' = np - q$.

▷ **Пример 2.2.** По данным примера **2.1** найти наивероятнейшее число появления бракованных деталей из 5 отобранных и вероятность этого числа.

Р е ш е н и е. По формуле **(2.4)** $5\cdot 0{,}2 - 0{,}8 \leq m_0 \leq 5\cdot 0{,}2 + 0{,}2$ или $0{,}2 \leq m_0 \leq 1{,}2$. Единственное целое число, удовлетворяющее полученному неравенству, $m_0 = 1$, а его вероятность $P_{1,5} = 0{,}4096$ была получена в примере **2.1**. ▶

▷ **Пример 2.3.** Сколько раз необходимо подбросить игральную кость, чтобы наивероятнейшее выпадение тройки было равно 10?

Р е ш е н и е. В данном случае $p = \frac{1}{6}$. Согласно неравенству **(2.4)** $n\cdot\frac{1}{6} - \frac{5}{6} \leq 10 \leq n\cdot\frac{1}{6} + \frac{1}{6}$ или $n - 5 \leq 60 \leq n + 1$, откуда $59 \leq n \leq 65$, т.е. необходимо подбросить кость от 59 до 65 раз (включительно). ▶

2.2. Формула Пуассона

Предположим, что мы хотим вычислить вероятность $P_{m,n}$ появления события A при большом числе испытаний n, например, $P_{300,500}$. По формуле Бернулли **(2.1)**

$$P_{300,500} = C_{500}^{300}p^{300}q^{200} = \frac{500!}{300!\,200!}p^{300}q^{200}.$$

Ясно, что в этом случае непосредственное вычисление по формуле Бернулли технически сложно, тем более если учесть, что сами p и q — числа дробные. Поэтому возникает естественное желание иметь более простые приближенные формулы для вычисления $P_{m,n}$ при больших n. Такие формулы, называемые *асимптотическими*, существуют и определяются теоремой Пуассона, локальной и интегральной теоремами Муавра—Лапласа. Наиболее простой из них является теорема Пуассона.

Теорема. *Если вероятность p наступления события A в каждом испытании стремится к нулю* ($p \to 0$) *при неограниченном увеличении числа n испытаний* ($n \to \infty$), *причем произведение np стре-*

мится к постоянному числу $\lambda(np \to \lambda)$, *то вероятность* $P_{m,n}$ *того, что событие* A *появится* m *раз в* n *независимых испытаниях, удовлетворяет предельному равенству*

$$\lim_{n\to\infty} P_{m,n} = P_m(\lambda) = \frac{\lambda^m e^{-\lambda}}{m!}. \qquad (2.5)$$

□ По формуле Бернулли **(2.1)**

$$P_{m,n} = C_n^m p^m q^{n-m} = \frac{n(n-1)(n-2)\ldots(n-m+1)}{m!} p^m (1-p)^n (1-p)^{-m}$$

или, учитывая, что $\lim\limits_{n\to\infty} np = \lambda$, т.е. при достаточно больших n

$$p \approx \frac{\lambda}{n} \text{ и } P_{m,n} \approx \frac{\lambda^m}{m!}\left(1\cdot\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{m-1}{n}\right)\right)\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-m}$$

Так как $\lim\limits_{n\to\infty}\left(1-\frac{1}{n}\right) = \lim\limits_{n\to\infty}\left(1-\frac{2}{n}\right) = \ldots = \lim\limits_{n\to\infty}\left(1-\frac{m-1}{n}\right) = 1,$

$\lim\limits_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = \lim\limits_{n\to\infty}\left(\left(1-\frac{\lambda}{n}\right)^{\frac{-n}{\lambda}}\right)^{-\lambda} = e^{-\lambda}$ и $\lim\limits_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-m} = 1,$ то

$\lim\limits_{n\to\infty} P_{m,n} = \frac{\lambda^m}{m!} e^{-\lambda}$. ■

Строго говоря, условие теоремы Пуассона $p \to 0$ при $n \to \infty$, так что $np \to \lambda$, противоречит исходной предпосылке схемы испытаний Бернулли, согласно которой вероятность наступления события в каждом испытании p = const. Однако, если вероятность p — постоянна и мала, число испытаний n — велико и число $\lambda = np$ — незначительно (будем полагать, что $\lambda = np \le 10$), то из предельного равенства **(2.5)** вытекает приближенная *формула Пуассона*:

$$P_{m,n} \approx \frac{\lambda^m e^{-\lambda}}{m!} = P_m(\lambda). \qquad (2.6)$$

В табл. III приложений приведены значения функции Пуассона $P_m(\lambda)$.

▷ **Пример 2.4.** На факультете насчитывается 1825 студентов. Какова вероятность того, что 1 сентября является днем рождения одновременно четырех студентов факультета?

Р е ш е н и е. Вероятность того, что день рождения студента 1 сентября, равна $p = 1/365$. Так как $p = 1/365$ — мала, $n = 1825$ — велико и $\lambda = np = 1825 \cdot (1/365) = 5 \le 10$, то применяем формулу Пуассона **(2.6)**:

$P_{4,1825} = P_4(5) = 0{,}1755$ (по табл. III приложений). ▶

2.3. Локальная и интегральная формулы Муавра—Лапласа

Локальная теорема Муавра—Лапласа. *Если вероятность* p *наступления события* A *в каждом испытании постоянна и отлична от* 0 *и* 1, *то вероятность* $P_{m,n}$ *того, что событие* A *произойдет* m *раз в* n *независимых испытаниях при достаточно большом числе* n, *приближенно равна*[1]

$$P_{m,n} \approx \frac{f(x)}{\sqrt{npq}}, \qquad (2.7)$$

где

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} — \qquad (2.8)$$

функция Гаусса

и

$$x = \frac{m-np}{\sqrt{npq}}. \qquad (2.9)$$

Чем больше n, тем точнее приближенная формула **(2.7)**, называемая *локальной формулой Муавра—Лапласа*. Приближенные значения вероятности $P_{m,n}$, даваемые локальной формулой **(2.7)**, на практике используются как точные при npq порядка двух и более десятков, т.е. при условии $npq \ge 20$.

Для упрощения расчетов, связанных с применением формулы **(2.7)**, составлена таблица значений функции $f(x)$ (табл. I, приведенная в приложении). Пользуясь этой таблицей, необходимо иметь в виду очевидные с в о й с т в а функции $f(x)$ **(2.8)**.

1. *Функция* $f(x)$ *является четной*, т.е. $f(-x) = f(x)$.

2. *Функция* $f(x)$ *— монотонно убывающая при положительных значениях* x, *причем при* $x \to \infty$ $f(x) \to 0$.

(Практически можно считать, что уже при $x > 4$ $f(x) \approx 0$.)

[1] Доказательство теоремы приведено в § 6.5. Вероятностный смысл величин np, npq устанавливается в § 4.1 (см. замечание на с. 146—147).

▷ **Пример 2.5.** В некоторой местности из каждых 100 семей 80 имеют холодильники. Найти вероятность того, что из 400 семей 300 имеют холодильники.

Р е ш е н и е. Вероятность того, что семья имеет холодильник, равна $p = 80/100 = 0{,}8$. Так как $n = 100$ достаточно велико (условие $npq = 100 \cdot 0{,}8(1-0{,}8)=64 \geq 20$ выполнено), то применяем локальную формулу Муавра—Лапласа.

Вначале определим по **(2.9)** $x = \dfrac{300 - 400 \cdot 0{,}8}{\sqrt{400 \cdot 0{,}8 \cdot 0{,}2}} = -2{,}50.$

Тогда по формуле **(2.7)** $P_{300,400} \approx \dfrac{f(-2{,}50)}{\sqrt{100 \cdot 0{,}8 \cdot 0{,}2}} = \dfrac{f(2{,}50)}{\sqrt{64}} =$

$$= \frac{0{,}0175}{8} \approx 0{,}0022$$

(значение $f(2{,}50)$ найдено по табл. I приложений). Весьма малое значение вероятности $P_{300,400}$ не должно вызывать сомнения, так как кроме события «ровно 300 семей из 400 имеют холодильники» возможно еще 400 событий: «0 из 400», «1 из 400»,..., «400 из 400» со своими вероятностями. Все вместе эти события образуют полную группу, а значит, сумма их вероятностей равна единице. ▶

Пусть в условиях примера **2.5** необходимо найти вероятность того, что от 300 до 360 семей (включительно) имеют холодильники. В этом случае по теореме сложения вероятность искомого события

$$P_{400}(300 \leq m \leq 360) = P_{300,400} + P_{301,400} + \ldots + P_{360,400}.$$

В принципе вычислить каждое слагаемое можно по локальной формуле Муавра—Лапласа, но большое количество слагаемых делает расчет весьма громоздким. В таких случаях используется следующая теорема.

Интегральная теорема Муавра—Лапласа. *Если вероятность p наступления события A в каждом испытании постоянна и отлична от* 0 *и* 1, *то вероятность того, что число m наступления события A в n независимых испытаниях заключено в пределах от a до b (включительно), при достаточно большом числе n приближенно равна*

$$P_n(a \leq m \leq b) \approx \frac{1}{2}\left[\Phi(x_2) - \Phi(x_1)\right], \qquad \textbf{(2.10)}$$

где $$\Phi(x) = \frac{2}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt \ — \qquad \textbf{(2.11)}$$

функция (или *интеграл вероятностей*) *Лапласа*;

$$x_1 = \frac{a - np}{\sqrt{npq}}, \quad x_2 = \frac{b - np}{\sqrt{npq}}. \qquad \textbf{(2.12)}$$

(Доказательство теоремы приведено в § 6.5.)

Формула **(2.10)** называется *интегральной формулой Муавра—Лапласа*. Чем больше n, тем точнее эта формула. При выполнении условия $npq \geq 20$ интегральная формула **(2.10)**, так же как и локальная, дает, как правило, удовлетворительную для практики погрешность вычисления вероятностей.

Функция $\Phi(x)$ табулирована (см. табл. II приложений). Для применения этой таблицы нужно знать с в о й с т в а функции $\Phi(x)$

1. *Функция* $\Phi(x)$ *нечетная*, т.е. $\Phi(-x) = -\Phi(x)$.

□ $\Phi(-x) = \dfrac{2}{\sqrt{2\pi}} \int_0^{-x} e^{-t^2/2} dt$. Сделаем замену переменной $t = -z$. Тогда $dt = -dz$. Пределами интегрирования по переменной z будут 0 и x. Получим

$$\Phi(-x) = -\frac{2}{2\pi} \int_0^x e^{-z^2/2} dz = -\Phi(x),$$

поскольку величина определенного интеграла не зависит от обозначения переменной интегрирования. ■

2. *Функция* $\Phi(x)$ *монотонно возрастающая, причем при* $x \to +\infty$ $\Phi(x) \to 1$ (практически можно считать, что уже при $x > 4$ $\Phi(x) \approx 1$).

□ Так как производная интеграла по переменному верхнему пределу равна подынтегральной функции при значении верхнего предела, т.е. $\Phi'(x) = \dfrac{2}{\sqrt{2\pi}} e^{-x^2/2}$, и всегда положительна, то $\Phi(x)$ монотонно возрастает на всей числовой прямой.

Найдем

$$\lim_{x \to +\infty} \Phi(x) = \lim_{x \to +\infty} \frac{2}{2\pi} \int_0^{\lambda} e^{-t^2/2} dt = \frac{2}{2\pi} \int_0^{+\infty} e^{-t^2/2} dt.$$

Сделаем замену переменной $z = t/\sqrt{2}$, тогда $\sqrt{2}dz = dt$, пределы интегрирования не меняются, и

$$\lim_{x\to+\infty}\Phi(x)=\frac{2}{\sqrt{2\pi}}\int_0^{+\infty}e^{-z^2}\sqrt{2}dz=\frac{2}{\sqrt{\pi}}\cdot\frac{1}{2}\int_{-\infty}^{+\infty}e^{-z^2}dz,$$

(так как интеграл от четной функции

$$\int_{-\infty}^{+\infty}e^{-z^2}dz=2\int_0^{+\infty}e^{-z^2}dz\,).$$

Учитывая, что $\int_{-\infty}^{+\infty}e^{-z^2}dz=\sqrt{\pi}$ (*интеграл Эйлера—Пуассона*), получим

$$\lim_{x\to+\infty}\Phi(x)=\frac{2}{\sqrt{\pi}}\cdot\frac{1}{2}\cdot\sqrt{\pi}=1.\ \blacksquare$$

▷ **Пример 2.6.** По данным примера **2.5** вычислить вероятность того, что от 300 до 360 (включительно) семей из 400 имеют холодильники.

Р е ш е н и е. Применяем интегральную теорему Муавра—Лапласа ($npq = 64 \geq 20$). Вначале определим по **(2.12)**

$$x_1=\frac{300-400\cdot0{,}8}{\sqrt{400\cdot0{,}8\cdot0{,}2}}=-2{,}50,\qquad x_2=\frac{360-400\cdot0{,}8}{\sqrt{400\cdot0{,}8\cdot0{,}2}}=5{,}0.$$

Теперь по формуле **(2.10)**, учитывая свойства $\Phi(x)$, получим

$$P_{400}(300\le m\le360)\approx\frac{1}{2}[\Phi(5{,}0)-\Phi(-2{,}50)]=\frac{1}{2}[\Phi(5{,}0)+\Phi(2{,}50)]\approx$$

$$\approx\frac{1}{2}(1+0{,}9876)=0{,}9938$$

(по табл. II приложений $\Phi(2{,}50)=0{,}9876$, $\Phi(5{,}0)\approx1$). ▶

Рассмотрим с л е д с т в и е интегральной теоремы Муавра—Лапласа.

Следствие. *Если вероятность p наступления события A в каждом испытании постоянна и отлична от* 0 *и* 1, *то при достаточно большом числе n независимых испытаний вероятность того, что:*

а) число m наступлений события A отличается от произведения np не более, чем на величину $\varepsilon > 0$ *(по абсолютной величине), т.е.*

$$P_n(|m-np|\le\varepsilon)\approx\Phi\left(\frac{\varepsilon}{\sqrt{npq}}\right);\qquad\textbf{(2.13)}$$

б) *частость* $\frac{m}{n}$ *события A заключена в пределах* от α до β *(включительно)*[1], т.е.

$$P_n\left(\alpha\le\frac{m}{n}\le\beta\right)\approx\frac{1}{2}[\Phi(z_2)-\Phi(z_1)],\qquad\textbf{(2.14)}$$

где $$z_1=\frac{\alpha-p}{\sqrt{pq/n}},\qquad z_2=\frac{\beta-p}{\sqrt{pq/n}}.\qquad\textbf{(2.15)}$$

в) *частость* $\frac{m}{n}$ *события A отличается от его вероятности p не более, чем на величину* $\Delta > 0$ *(по абсолютной величине), т.е.*

$$P_n\left(\left|\frac{m}{n}-p\right|\le\Delta\right)\approx\Phi\left(\frac{\Delta\sqrt{n}}{\sqrt{pq}}\right).\qquad\textbf{(2.16)}$$

□ а) Неравенство $|m-np|\le\varepsilon$ равносильно двойному неравенству $np-\varepsilon\le m\le np+\varepsilon$. Поэтому по интегральной формуле **(2.10)**

$$P_n(|m-np|\le\varepsilon)=P_n(np-\varepsilon\le m\le np+\varepsilon)\approx$$

$$\approx\frac{1}{2}\left[\Phi\left(\frac{p+\varepsilon-np}{\sqrt{npqn}}\right)-\Phi\left(\frac{np-\varepsilon-np}{\sqrt{npq}}\right)\right]=\frac{1}{2}\left[\Phi\left(\frac{\varepsilon}{\sqrt{npq}}\right)-\Phi\left(\frac{-\varepsilon}{\sqrt{npq}}\right)\right]=$$

$$=\frac{1}{2}\left[\Phi\left(\frac{\varepsilon}{\sqrt{npq}}\right)+\Phi\left(\frac{\varepsilon}{\sqrt{npq}}\right)\right]=\Phi\left(\frac{\varepsilon}{\sqrt{npq}}\right).$$

б) Неравенство $\alpha\le\frac{m}{n}\le\beta$ равносильно неравенству $a\le m\le b$ при $a=n\alpha$ и $b=n\beta$. Заменяя в формулах **(2.10)**, **(2.12)** величины a и b полученными выражениями, получим доказываемые формулы **(2.14)** и **(2.15)**.

в) Неравенство $\left|\frac{m}{n}-p\right|\le\Delta$ равносильно неравенству $|m-np|\le\Delta n$. Заменяя в формуле **(2.13)** $\varepsilon=\Delta n$, получим доказываемую формулу **(2.16)**. ■

[1] Вероятностный смысл величины pq/n устанавливается в § 4.1.

▷ **Пример 2.7.** По данным примера **2.5** вычислить вероятность того, что от 280 до 360 семей из 400 имеют холодильники

Р е ш е н и е. Вычислить вероятность $P_{400}(280 \le m \le 360)$ можно аналогично примеру **2.6** по основной формуле (**2.10**). Но проще это сделать, если заметить, что границы интервала 280 и 360 симметричны относительно величины $np = 320$. Тогда по формуле (**2.13**)

$$P_{400}(280 \le m \le 360) = P_{400}(-40 \le m - 320 \le 40) =$$

$$= P_{400}(|m - 320| \le 40) \approx \Phi\left(\frac{40}{\sqrt{400 \cdot 0{,}8 \cdot 0{,}2}}\right) = \Phi(5{,}0) \approx 1. \blacktriangleright$$

▷ **Пример 2.8.** По статистическим данным в среднем 87% новорожденных доживают до 50 лет. 1. Найти вероятность того, что из 1000 новорожденных доля (частость) доживших до 50 лет будет: а) заключена в пределах от 0,9 до 0,95; б) будет отличаться от вероятности этого события не более, чем на 0,04 (по абсолютной величине). 2. При каком числе новорожденных с надежностью 0,95 доля доживших до 50 лет будет заключена в границах от 0,86 до 0,88?

Р е ш е н и е. 1. а) Вероятность p того, что новорожденный доживет до 50 лет, равна 0,87. Так как $n = 1000$ велико (условие $npq = 1000 \cdot 0{,}87 \cdot 0{,}13 = 113{,}1 \ge 20$ выполнено), то используем следствие интегральной теоремы Муавра—Лапласа. Вначале определим по (**2.15**)

$$z_1 = \frac{0{,}9 - 0{,}87}{\sqrt{0{,}87 \cdot 0{,}13/1000}} = 2{,}82, \quad z_2 = \frac{0{,}95 - 0{,}87}{\sqrt{0{,}87 \cdot 0{,}13/1000}} = 7{,}52.$$

Теперь по формуле (**2.14**)

$$P_{1000}\left(0{,}9 \le \frac{m}{n} \le 0{,}95\right) \approx \frac{1}{2}[\Phi(7{,}52) - \Phi(2{,}82)] =$$

$$= \frac{1}{2}(1 - 0{,}9952) = 0{,}0024.$$

б) По формуле (**2.16**)

$$P_{1000}\left(\left|\frac{m}{n} - 0{,}87\right| \le 0{,}04\right) \approx \Phi\left(\frac{0{,}04 \cdot \sqrt{1000}}{\sqrt{0{,}87 \cdot 0{,}13}}\right) = \Phi(3{,}76) = 0{,}9998.$$

Так как неравенство $\left|\frac{m}{n} - 0{,}87\right| \le 0{,}04$ равносильно неравенству $0{,}83 \le \frac{m}{n} \le 0{,}91$, полученный результат означает, что практически достоверно, что от 0,83 до 0,91 числа новорожденных из 1000 доживут до 50 лет. ▶

2. По условию $P_n(0{,}86 \le \frac{m}{n} \le 0{,}88) = 0{,}95$, или $P_n(-0{,}01 \le \frac{m}{n} - 0{,}87 \le 0{,}01) = P_n\left[\left|\frac{m}{n} - 0{,}87\right| \le 0{,}01\right] = 0{,}95.$ (*)

По формуле (**2.16**) при $\Delta = 0{,}01$ $\Phi\left(\frac{\Delta\sqrt{n}}{\sqrt{pq}}\right) = 0{,}95.$

По табл. II приложений $\Phi(t) = 0{,}95$ при $t = 1{,}96$, следовательно, $\frac{\Delta\sqrt{n}}{\sqrt{pq}} = t$, откуда

$$n = \frac{t^2 pq}{\Delta^2} = \frac{1{,}96^2 \cdot 0{,}87 \cdot 0{,}13}{0{,}01^2} = 4345,$$

т.е. условие (*) может быть гарантировано при существенном увеличении числа рассматриваемых новорожденных до $n = 4345$. ▶

2.4. Решение задач

▷ **Пример 2.9.** В среднем 20% пакетов акций на аукционах продаются по первоначально заявленной цене. Найти вероятность того, что из 9 пакетов акций в результате торгов по первоначально заявленной цене: 1) не будут проданы 5 пакетов; 2) будет продано: а) менее 2 пакетов; б) не более 2; в) хотя бы 2 пакета; г) наивероятнейшее число пакетов.

Р е ш е н и е. 1) Вероятность того, что пакет акций не будет продан по первоначально заявленной цене, $p = 1 - 0{,}2 = 0{,}8$.

По формуле Бернулли (**2.1**)

$$P_{5,9} = C_9^5 \cdot 0{,}8^5 \cdot 0{,}2^4 = 0{,}066.$$

2.а) По условию $p = 0{,}2$.

$$P_9(m < 2) = P_{0,9} + P_{1,9} = C_9^0 \cdot 0{,}2^0 \cdot 0{,}8^9 + C_9^1 \cdot 0{,}2 \cdot 0{,}8^8 = 0{,}436.$$

2.б) $P_9(m \le 2) = P_{0,9} + P_{1,9} + P_{2,9} = C_9^0 \cdot 0{,}2^0 \cdot 0{,}8^9 + C_9^1 \cdot 0{,}2 \cdot 0{,}8^8 +$

$$+ C_9^2 \cdot 0{,}2^2 \cdot 0{,}8^7 = 0{,}738.$$

2.в) $P_9(m \geq 2) = P_{2,9} + P_{3,9} + \ldots + P_{9,9}$.

Указанную вероятность можно найти проще, если перейти к противоположному событию, т.е.

$P_9(m \geq 2) = 1 - P_9(m < 2) = 1 - (P_{0,9} + P_{1,9}) = 1 - 0{,}436 = 0{,}564$ (см. п. 2а)).

2.г) Наивероятнейшее число проданных акций по первоначально заявленной цене определится из условия (**2.4**), т.е.

$9 \cdot 0{,}2 - 0{,}8 \leq m_0 \leq 9 \cdot 0{,}2 + 0{,}2$ или $1 \leq m_0 \leq 2$, т.е. наивероятнейших чисел два: m_0=1 и $m_0' = 2$. Поэтому вероятность

$$P_{\text{наивер}} = P_{1,9} + P_{2,9} = C_9^1 \cdot 0{,}2 \cdot 0{,}8^8 + C_9^2 \cdot 0{,}2^2 \cdot 0{,}8^7 = 0{,}604. \blacktriangleright$$

▷ **Пример 2.10.** Завод отправил на базу 10 000 стандартных изделий. Среднее число изделий, повреждаемых при транспортировке, составляет 0,02%. Найти вероятность того, что из 10 000 изделий: 1) будет повреждено: а) 3; б) по крайней мере 3; 2) не будет повреждено: а) 9997; б) хотя бы 9997.

Р е ш е н и е. 1. а) Вероятность того, что изделие будет повреждено при транспортировке, равна $p = 0{,}0002$. Так как p — мала, $n = 10\,000$ — велико и $\lambda = np = 10\,000 \cdot 0{,}0002 = 2 \leq 10$, следует применить формулу Пуассона (**2.6**): $P_{3,10\,000} = \dfrac{2^3 e^{-2}}{3!}$.

Это значение проще найти, используя табл. III приложений:

$$P_{3,10\,000} = P_3(2) = 0{,}1804.$$

1.б) Вероятность $P_{10\,000}(m \geq 3)$ может быть вычислена как сумма большого количества слагаемых:

$$P_{10\,000}(m \geq 3) = P_{3,10\,000} + P_{4,10\,000} + \ldots + P_{10\,000,10\,000}.$$

Но, разумеется, проще ее найти, перейдя к противоположному событию:

$$P_{10\,000}(m \geq 3) = 1 - P_{10\,000}(m < 3) = 1 - (P_{0,10000} + P_{1,10000} + P_{2,10000}) =$$

$$= 1 - (0{,}1353 + 0{,}2707 + 0{,}2707) = 0{,}3233.$$

Следует отметить, что для вычисления вероятности $P_{10\,000}(m \geq 3) = P_{10\,000}(3 \leq m \leq 10\,000)$ нельзя применить интегральную формулу

Муавра—Лапласа, так как не выполнено условие ее применимости, ибо $npq \approx 2 < 20$.

2.а) В данном случае $p = 1 - 0{,}0002 = 0{,}9998$ и надо найти $P_{9997,10\,000}$, для непосредственного вычисления которой нельзя применить ни формулу Пуассона (p велика), ни локальную формулу Муавра—Лапласа $(npq \approx 2 < 20)$. Однако событие «не будет повреждено 9997 из 10 000» равносильно событию «будет повреждено 3 из 10 000», вероятность которого, равная 0,1804, получена в п. а).

2.б) Событие «не будет повреждено хотя бы 9997 из 10 000» равносильно событию «будет повреждено не более 3 из 10 000», для которого $p = 0{,}0002$ и

$$P_{10\,000}(m \leq 3) = P_{0,10\,000} + P_{1,10\,000} + P_{2,10\,000} + P_{3,10\,000} =$$

$$= 0{,}1353 + 0{,}2707 + 0{,}2707 + 0{,}1805 = 0{,}8572. \blacktriangleright$$

▷ **Пример 2.11.** По результатам проверок налоговыми инспекциями установлено, что в среднем каждое второе малое предприятие региона имеет нарушение финансовой дисциплины. Найти вероятность того, что из 1000 зарегистрированных в регионе малых предприятий имеют нарушения финансовой дисциплины: а) 480 предприятий; б) наивероятнейшее число предприятий; в) не менее 480; г) от 480 до 520.

Р е ш е н и е. а) По условию $p = 0{,}5$. Так $n = 1000$ достаточно велико (условие $npq = 10\,000 \cdot 0{,}5(1 - 0{,}5) = 250 \geq 20$ выполнено), то применяем локальную формулу Муавра—Лапласа. Вначале по (**2.9**) определим $x = \dfrac{480 - 1000 \cdot 0{,}5}{\sqrt{1000 \cdot 0{,}5 \cdot 0{,}5}} = -1{,}265$, затем по формуле (**2.7**)[1]

[1] При вычислении значений $f(1{,}265)$ и $\Phi(1{,}265)$ используем линейную интерполяцию (см. табл. I и II приложений):

$$f(1{,}265) \approx \frac{f(1{,}26) + f(1{,}27)}{2} = \frac{1}{2}(0{,}1804 + 0{,}1781) = 0{,}1792,$$

$$\Phi(1{,}265) \approx \frac{\Phi(1{,}26) + \Phi(1{,}27)}{2} = \frac{1}{2}(0{,}7923 + 0{,}7959) = 0{,}7941.$$

$$P_{480,1000} \approx \frac{f(-1{,}265)}{\sqrt{1000 \cdot 0{,}5 \cdot 0{,}5}} = \frac{f(1{,}265)}{\sqrt{250}} = \frac{0{,}1792}{\sqrt{250}} = 0{,}0113.$$

б) По формуле **(2.6)** наивероятнейшее число $1000 \cdot 0{,}5 - 0{,}5 \le m_0 \le 1000 \cdot 0{,}5 + 0{,}5$, т.е. $499{,}5 \le m_0 \le 500{,}5$ и целое $m_0 = 500$. Теперь по **(2.9)** определим

$$x = \frac{500 - 1000 \cdot 0{,}5}{\sqrt{1000 \cdot 0{,}5 \cdot 0{,}5}} = 0 \text{ и } P_{500,1000} \approx \frac{f(0)}{\sqrt{250}} = \frac{0{,}3989}{\sqrt{250}} = 0{,}0252.$$

в) Необходимо найти

$P_{1000}(m \ge 480) = P_{1000}(480 \le m \le 1000)$. Применяем интегральную формулу Муавра—Лапласа **(2.10)**, предварительно найдя по формуле **(2.12)**

$$x_1 = \frac{480 - 1000 \cdot 0{,}5}{\sqrt{1000 \cdot 0{,}5 \cdot 0{,}5}} = -1{,}265, \quad x_2 = \frac{1000 - 1000 \cdot 0{,}5}{\sqrt{1000 \cdot 0{,}5 \cdot 0{,}5}} = 31{,}6.$$

Теперь

$$P_{1000}(480 \le m \le 1000) \approx \frac{1}{2}[\Phi(31{,}6) - \Phi(-1{,}265)] =$$

$$= \frac{1}{2}[\Phi(31{,}6) + \Phi(1{,}265)] \approx \frac{1}{2}(1 + 0{,}7941) \approx 0{,}897.$$

г) Вероятность $P_{1000}(480 \le m \le 520)$ можно было найти по той же интегральной формуле Муавра—Лапласа **(2.10)**. Но проще это сделать, используя следствие **(2.13)**, заметив, что границы интервала 480 и 520 симметричны относительно значения $np = 1000 \cdot 0{,}5 = 500$:

$$P_{1000}(480 \le m \le 520) = P_{1000}(|m - 500| \le 20) \approx \Phi\left(\frac{20}{\sqrt{250}}\right) =$$

$$= \Phi(1{,}265) = 0{,}794. \blacktriangleright$$

▷ **Пример 2.12.** В страховой компании 10 тыс. клиентов. Страховой взнос каждого клиента составляет 500 руб. При наступлении страхового случая, вероятность которого по имеющимся данным и оценкам экспертов можно считать равной $p = 0{,}005$, страховая компания обязана выплатить клиенту страховую сумму размером 50 тыс. руб. На какую прибыль может рассчитывать страховая компания с надежностью 0,95?

Р е ш е н и е. Размер прибыли компании составляет разность между суммарным взносом всех клиентов и суммарной страховой суммой, выплаченной n_0 клиентам при наступлении страхового случая, т.е.

$$П = 500 \cdot 10 - 50n_0 = 50(100 - n_0) \text{ тыс. руб.}$$

Для определения n_0 применим интегральную формулу Муавра—Лапласа (требование $npq = 10\,000 \cdot 0{,}005 \cdot 0{,}995 = 49{,}75 \ge 20$ выполнено).

По условию задачи

$$P_{10000}(0 \le m \le n_0) = \frac{1}{2}[\Phi(x_2) - \Phi(x_1)] = 0{,}95, \qquad \textbf{(2.17)}$$

где m — число клиентов, которым будет выплачена страховая сумма;

$$x_1 = \frac{0 - np}{\sqrt{npq}} = -\sqrt{\frac{np}{q}} = -\sqrt{\frac{10\,000 \cdot 0{,}005}{0{,}995}} = -7{,}09, \quad x_2 = \frac{n_0 - np}{\sqrt{npq}},$$

откуда

$$n_0 = np + x_2\sqrt{npq} = 10\,000 \cdot 0{,}005 + x_2\sqrt{49{,}75} = 50 + x_2\sqrt{49{,}75}.$$

Из соотношения **(2.17)**

$$\Phi(x_2) = 1{,}9 + \Phi(x_1) = 1{,}9 + \Phi(-7{,}09) \approx 1{,}9 + (-1) = 0{,}9.$$

По табл. II приложений $\Phi(x_2) = 0{,}9$ при $x_2 = 1{,}645$.

Теперь $n_0 = 50 + 1{,}645\sqrt{49{,}75} = 61{,}6$ и $П = 50(100 - 61{,}6) = 1920$, т.е. с надежностью 0,95 ожидаемая прибыль составит 1,92 млн руб. ▶

2.5. Полиномиальная схема

Как отмечено выше, схема Бернелли представляет последовательность независимых испытаний с двумя исходами. При этом в каждом испытании событие A может появиться с одной и той же вероятностью p, а событие $\overline{A}$ — с вероятностью $q = 1 - p$.

В *полиномиальной* (*мультиномиальной*) *схеме* осуществляется переход от последовательности независимых испытаний с двумя исходами (A и $\bar{A}$) к последовательности независимых испытаний с k исключающими друг друга исходами A_1, A_2, ..., A_k. При этом в каждом испытании события A_1, A_2, ..., A_k наступают соответственно с вероятностями p_1, p_2, ..., p_k. Тогда вероятность $P_n(m_1, m_2, ..., m_k)$ того, что в n независимых испытаниях событие A_1 произойдет m_1 раз, A_2 — m_2, и т.д., событие A_k — m_k раз ($m_1 + m_2 + ... + m_k = n$), определится по формуле:

$$P_n(m_1, m_2, ..., m_k) = \frac{n!}{m_1! m_2! ... m_k!} p_1^{m_1} p_2^{m_2} ... p_k^{m_k}. \qquad (2.18)$$

Формула (**2.18**) получается с учетом того, что событие, состоящее в появлении в n независимых испытаниях события A_1 m_1 раз, A_2 — m_2 и т.д., события A_k — m_k раз ($m_1 + m_2 + ... + m_k = n$), можно представить в виде суммы несовместных вариантов, вероятность каждого из которых по теореме умножения вероятностей для независимых событий равна $p_1^{m_1} p_2^{m_2} ... p_k^{m_k}$, а число вариантов определяется числом перестановок с повторениями (**1.15**) из n элементов.

В частном случае двух исходов при $m_1 = m$, $m_2 = n - m$, $p_1 = p$, $p_2 = q$, где $q = 1 - p$, формула (**2.18**) представляет формулу Бернулли (**2.1**).

▷ **Пример 2.12а.** Человек, принадлежащий к определенной группе населения, с вероятностью 0,2 оказывается брюнетом, с вероятностью 0,3 — шатеном, с вероятностью 0,4 — блондином и с вероятностью 0,1 — рыжим. Найти вероятность того, что в составе выбранной наудачу группы из 8 человек: а) равное число брюнетов, шатенов, блондинов и рыжих; б) число блондинов втрое больше числа рыжих.

Р е ш е н и е. а) По формуле (**2.18**) вероятность искомого события A равна

$$P(A) = P_8(2;2;2;2) = \frac{8!}{2!\,2!\,2!\,2!} 0{,}2^2 \cdot 0{,}3^2 \cdot 0{,}4^2 \cdot 0{,}1^2 = 0{,}0145.$$

б) Вероятность искомого события B равна сумме вероятностей двух несовместных событий (вариантов):

B_1 — в группе 3 блондина, 1 — рыжий, а остальные — ни то, ни другое;

B_2 — в группе 6 блондинов и 2 рыжих.

По формуле (**2.18**), полагая, что $p_1 = 0{,}4$; $p_2 = 0{,}1$; $p_3 = 1 - (0{,}4 + 0{,}1) = 0{,}5$, найдем

$$P(B_1) = P_8(3;1;4) = \frac{8!}{3!\,1!\,4!} 0{,}4^3 \cdot 0{,}1 \cdot 0{,}5^4 = 0{,}1120;$$

$$P(B_2) = P_8(6;2) = \frac{8!}{6!\,2!} 0{,}4^6 \cdot 0{,}1^2 = 0{,}0011;$$

$$P(B) = P(B_1) + P(B_2) = 0{,}1120 + 0{,}0011 = 0{,}1131. ◀$$

Если вероятности p_1, p_2,..., p_k наступления событий A_1, A_2,..., A_k в каждом испытании меняются в зависимости от исходов других, то мы имеем схему зависимых испытаний.

Последовательность зависимых испытаний, в которых условные вероятности наступления событий A_1, A_2,..., A_k в каждом (n + +1-ом) испытании (n = 1, 2,...) зависят только от исхода предшествующего (n-го) испытания, называются *цепями Маркова*. Цепи Маркова представляют один из видов марковского случайного процесса, рассматриваемого (в иной терминологии) в § 7.3.

Упражнения

2.13. Вероятность малому предприятию быть банкротом за время t равна 0,2. Найти вероятность того, что из шести малых предприятий за время t сохранятся: а) два; б) более двух.

2.14. В среднем пятая часть поступающих в продажу автомобилей некомплектны. Найти вероятность того, что среди десяти автомобилей имеют некомплектность: а) три автомобиля; б) менее трех.

2.15. Производится залп из шести орудий по некоторому объекту. Вероятность попадания в объект из каждого орудия равна 0,6. Найти вероятность ликвидации объекта, если для этого необходимо не менее четырех попаданий.

2.16. В среднем по 15% договоров страховая компания выплачивает страховую сумму. Найти вероятность того, что из десяти договоров с наступлением страхового случая будет связано с выплатой страховой суммы: а) три договора; б) менее двух договоров.

2.17. Предполагается, что 10% открывающихся новых малых предприятий прекращают свою деятельность в течение года. Какова вероятность того, что из шести малых

предприятий не более двух в течение года прекратят свою деятельность?

2.18. В семье десять детей. Считая вероятности рождения мальчика и девочки равными между собой, определить вероятность того, что в данной семье: а) не менее трех мальчиков; б) не более трех мальчиков.

2.19. Два равносильных противника играют в шахматы. Что более вероятно: а) выиграть 2 партии из 4 или 3 партии из 6; б) не менее 2 партий из 6 или не менее 3 партий из 6? (Ничьи в расчет не принимаются.)

2.20. В банк отправлено 4000 пакетов денежных знаков. Вероятность того, что пакет содержит недостаточное или избыточное число денежных знаков, равна 0,0001. Найти вероятность того, что при проверке будет обнаружено: а) три ошибочно укомплектованных пакета; б) не более трех пакетов.

2.21. Строительная фирма, занимающаяся установкой летних коттеджей, раскладывает рекламные листки по почтовым ящикам. Прежний опыт работы компании показывает, что примерно в одном случае из двух тысяч следует заказ. Найти вероятность того, что при размещении 100 тыс. листков число заказов будет: а) равно 48; б) находиться в границах от 45 до 55.

2.22. В вузе обучаются 3650 студентов. Вероятность того, что день рождения студента приходится на определенный день года, равна 1/365. Найти: а) наиболее вероятное число студентов, родившихся 1 мая, и вероятность такого события; б) вероятность того, что по крайней мере 3 студента имеют один и тот же день рождения.

2.23. Учебник издан тиражом 10 000 экземпляров. Вероятность того, что экземпляр учебника сброшюрован неправильно, равна 0.0001. Найти вероятность того, что: а) тираж содержит 5 бракованных книг; б) по крайней мере 9998 книг сброшюрованы правильно.

2.24. Два баскетболиста делают по 3 броска мячом в корзину. Вероятности попадания мяча в корзину при каждом броске равны соответственно 0,6 и 0,7. Найти вероятность того, что: а) у обоих будет одинаковое количество попаданий; б) у первого баскетболиста будет больше попаданий, чем у второго.

2.25. Известно, что в среднем 60% всего числа изготовляемых заводом телефонных аппаратов является продукцией первого сорта. Чему равна вероятность того, что в изготовленной партии окажется: а) 6 аппаратов первого сорта, если партия содержит 10 аппаратов; б) 120 аппаратов первого сорта, если партия содержит 200 аппаратов?

2.26. Вероятность того, что перфокарта набита оператором неверно, равна 0,1. Найти вероятность того, что: а) из 200 перфокарт правильно набитых будет не меньше 180; б) у того же оператора из десяти перфокарт будет неверно набитых не более двух.

2.27. Аудиторную работу по теории вероятностей с первого раза успешно выполняют 50% студентов. Найти вероятность того, что из 400 студентов работу успешно выполнят: а) 180 студентов, б) не менее 180 студентов.

2.28. При обследовании уставных фондов банков установлено, что пятая часть банков имеют уставный фонд свыше 100 млн руб. Найти вероятность того, что среди 1800 банков имеют уставный фонд свыше 100 млн руб.: а) не менее 300; б) от 300 до 400 включительно.

2.29. Сколько нужно взять деталей, чтобы наивероятнейшее число годных деталей было равно 50, если вероятность того, что наудачу взятая деталь будет бракованной, равна 0,1?

2.30. Вероятность того, что пассажир опоздает к отправлению поезда, равна 0,01. Найти наиболее вероятное число опоздавших из 800 пассажиров и вероятность такого числа опоздавших.

2.31. Вероятность того, что деталь стандартна, равна $p = 0{,}9$. Найти: а) с вероятностью 0,9545 границы (симметричные относительно p), в которых заключена доля стандартных среди проверенных 900 деталей; б) вероятность того, что доля нестандартных деталей среди них заключена в пределах от 0,08 до 0,11.

2.32. В результате проверки качества приготовленных для посева семян гороха установлено, что в среднем 90% всхожи. Сколько нужно посеять семян, чтобы с вероятностью 0,991 можно было ожидать, что доля взошедших семян отклонится от вероятности взойти каждому семени не более, чем на 0,03 (по абсолютной величине)?

Упражнения

3.25. Вероятность поражения вирусным заболеванием куста земляники равна 0,2. Составить закон распределения числа кустов земляники, зараженных вирусом, из четырех посаженных кустов.

3.26. Стрелок ведет стрельбу по цели с вероятностью попадания при каждом выстреле 0,2. За каждое попадание он получает 5 очков, а в случае промаха очков ему не начисляют. Составить закон распределения числа очков, полученных стрелком за 3 выстрела, и вычислить математическое ожидание этой случайной величины.

3.27. В рекламных целях торговая фирма вкладывает в каждую десятую единицу товара денежный приз размером 1 тыс. руб. Составить закон распределения случайной величины — размера выигрыша при пяти сделанных покупках. Найти математическое ожидание и дисперсию этой случайной величины.

3.28. Клиенты банка, не связанные друг с другом, не возвращают кредиты в срок с вероятностью 0,1. Составить закон распределения числа возвращенных в срок кредитов из 5 выданных. Найти математическое ожидание, дисперсию и среднее квадратическое отклонение этой случайной величины.

3.29. Контрольная работа состоит из трех вопросов. На каждый вопрос приведено 4 ответа, один из которых правильный. Составить закон распределения числа правильных ответов при простом угадывании. Найти математическое ожидание и дисперсию этой случайной величины.

3.30. В среднем по 10% договоров страховая компания выплачивает страховые суммы в связи с наступлением страхового случая. Составить закон распределения числа таких договоров среди наудачу выбранных четырех. Вычислить математическое ожидание и дисперсию этой случайной величины.

3.31. В билете три задачи. Вероятность правильного решения первой задачи равна 0,9, второй — 0,8, третьей — 0,7. Составить закон распределения числа правильно решенных задач в билете и вычислить математическое ожидание и дисперсию этой случайной величины.

3.32. Вероятность попадания в цель при одном выстреле равна 0,8 и уменьшается с каждым выстрелом на 0,1. Составить закон распределения числа попаданий в цель, если сделано три выстрела. Найти математическое ожидание, дисперсию и среднее квадратическое отклонение этой случайной величины.

3.33. Произведено два выстрела в мишень. Вероятность попадания в мишень первым стрелком равна 0,8, вторым — 0,7, Составить закон распределения числа попаданий в мишень. Найти математическое ожидание, дисперсию и функцию распределения этой случайной величины и построить ее график. (Каждый стрелок делает по одному выстрелу.)

3.34. Найти закон распределения числа пакетов трех акций, по которым владельцем будет получен доход, если вероятность получения дохода по каждому из них равна соответственно 0,5, 0,6, 0,7. Найти математическое ожидание и дисперсию данной случайной величины, построить функцию распределения.

3.35. Дан ряд распределения случайной величины

X:	x_i	2	4
	p_i	p_1	p_2

Найти функцию распределения этой случайной величины, если ее математическое ожидание равно 3,4, а дисперсия равна 0,84.

3.36. Из пяти гвоздик две белые. Составить закон распределения и найти функцию распределения случайной величины, выражающей число белых гвоздик среди двух одновременно взятых.

3.37. Из 10 телевизоров на выставке 4 оказались фирмы «Сони». Наудачу для осмотра выбрано 3. Составить закон распределения числа телевизоров фирмы «Сони» среди 3 отобранных.

3.38. Среди 15 собранных агрегатов 6 нуждаются в дополнительной смазке. Составить закон распределения числа агрегатов, нуждающихся в дополнительной смазке, среди пяти наудачу отобранных из общего числа.

3.39. В магазине продаются 5 отечественных и 3 импортных телевизора. Составить закон распределения случайной

величины — числа импортных из четырех наудачу выбранных телевизоров. Найти функцию распределения этой случайной величины и построить ее график.

3.40. Вероятность того, что в библиотеке необходимая студенту книга свободна, равна 0,3. Составить закон распределения числа библиотек, которые посетит студент, если в городе 4 библиотеки. Найти математическое ожидание и дисперсию этой случайной величины.

3.41. Экзаменатор задает студенту вопросы, пока тот правильно отвечает. Как только число правильных ответов достигнет четырех либо студент ответит неправильно, экзаменатор прекращает задавать вопросы. Вероятность правильного ответа на один вопрос равна 2/3. Составить закон распределения числа заданных студенту вопросов.

3.42. Торговый агент имеет 5 телефонных номеров потенциальных покупателей и звонит им до тех пор, пока не получит заказ на покупку товара. Вероятность того, что потенциальный покупатель сделает заказ, равна 0,4. Составить закон распределения числа телефонных разговоров, которые предстоит провести агенту. Найти математическое ожидание и дисперсию этой случайной величины.

3.43. Каждый поступающий в институт должен сдать 3 экзамена. Вероятность успешной сдачи первого экзамена 0,9, второго — 0,8, третьего — 0,7. Следующий экзамен поступающий сдает только в случае успешной сдачи предыдущего. Составить закон распределения числа экзаменов, сдававшихся поступающим в институт. Найти математическое ожидание этой случайной величины.

3.44. Охотник, имеющий 4 патрона, стреляет по дичи до первого попадания или до израсходования всех патронов. Вероятность попадания при первом выстреле равна 0,6, при каждом последующем — уменьшается на 0,1. Необходимо: а) составить закон распределения числа патронов, израсходованных охотником; б) найти математическое ожидание и дисперсию этой случайной величины.

3.45. Из поступивших в ремонт 10 часов 7 нуждаются в общей чистке механизма. Часы не рассортированы по виду ремонта. Мастер, желая найти часы, нуждающиеся в чистке, рассматривает их поочередно и, найдя такие часы, прекращает дальнейший просмотр. Составить закон распределения числа просмотренных часов. Найти математическое ожидание и дисперсию этой случайной величины.

3.46. Имеются 4 ключа, из которых только один подходит к замку. Составить закон распределения числа попыток открывания замка, если испробованный ключ в последующих попытках не участвует. Найти математическое ожидание, дисперсию и среднее квадратическое отклонение этой случайной величины.

3.47. Абонент забыл последнюю цифру нужного ему номера телефона, однако помнит, что она нечетная. Составить закон распределения числа сделанных им наборов номера телефона до попадания на нужный номер, если последнюю цифру он набирает наудачу, а набранную цифру в дальнейшем не набирает. Найти математическое ожидание и функцию распределения этой случайной величины.

3.48. Дана функция распределения случайной величины X

$$F(x)=\begin{cases}0 & \text{при } x \le 1,\\ 0{,}3 & \text{при } 1 < x \le 2,\\ 0{,}7 & \text{при } 2 < x \le 3,\\ 1 & \text{при } x > 3.\end{cases}$$

Найти: а) ряд распределения; б) $M(X)$ и $D(X)$; в) построить многоугольник распределения и график $F(x)$.

3.49. Даны законы распределения двух независимых случайных величин

X:

x_i	0	1	3
p_i	0,2	0,5	?

и Y:

y_i	2	3
p_i	0,4	?

Найти вероятности, с которыми случайные величины принимают значение 3, а затем составить закон распределения случайной величины $3X-2Y$ и проверить выполнение свойств математических ожиданий и дисперсий:

$M(3X-2Y)=3M(X)-2M(Y)$, $D(3X-2Y)=9D(X)+4D(Y)$.

3.50. На двух автоматических станках производятся динаковые изделия. Даны законы распределения числа бракованных изделий, производимых в течение смены на каждом из них:

а) для первого

X:

x_i	0	1	2
p_i	0,1	0,6	0,3

,

б) для второго

Y:

y_j	0	2
p_j	0,5	0,5

Необходимо: а) составить закон распределения числа производимых в течение смены бракованных изделий обоими станками; б) проверить свойство математического ожидания суммы случайных величин.

3.51. Одна из случайных величин задана законом распределения

x_i	−1	0	1
p_i	0,1	0,8	0,1

,

а другая имеет биномиальное распределение с параметрами $n = 2$, $p = 0{,}6$. Составить закон распределения их суммы и найти математическое ожидание этой случайной величины.

3.52. Случайные величины X и Y независимы и имеют один и тот же закон распределения:

Значение	1	2	4
Вероятность	0,2	0,3	0,5

Составить закон распределения случайных величин $2X$ и $X+Y$. Убедиться в том, что $2X \neq X+Y$, но $M(2X) = M(X+Y)$.

3.53. По данным примера **3.52** убедиться в том, что $X^2 \neq XY$. Проверить равенство $M(XY) = [M(X)]^2$.

3.54. Два стрелка сделали по два выстрела по мишени. Вероятность попадания в мишень для первого стрелка равна 0,6, для второго — 0,7. Необходимо: а) составить закон распределения общего числа попаданий; б) найти математическое ожидание и дисперсию этой случайной величины.

3.55. Пусть X, Y, Z — случайные величины: X — выручка фирмы, Y — ее затраты, $Z = X-Y$ — прибыль. Найти распределение прибыли Z, если затраты и выручка независимы и заданы распределениями:

X:

x_i	3	4	5
p_i	1/3	1/3	1/3

,

Y:

y_j	1	2
p_j	1/2	1/2

3.56. Пусть X — выручка фирмы в долларах. Найти распределение выручки в рублях $Z=X\cdot Y$ в пересчете по курсу доллара Y, если выручка X не зависит от курса Y, а распределения X и Y имеют вид

X:

x_i	1000	2000
p_i	0,7	0,3

и Y:

y_j	25	27
p_j	0,4	0,6

3.57. Сделано два высокорисковых вклада: 10 тыс. руб. в компанию A и 15 тыс. руб. — в компанию B. Компания A обещает 50% годовых, но может «лопнуть» с вероятностью 0,2. Компания B обещает 40% годовых, но может «лопнуть» с вероятностью 0,15. Составить закон распределения случайной величины — общей суммы прибыли (убытка), полученной от двух компаний через год, и найти ее математическое ожидание.

3.58. Дискретная случайная величина X задана рядом распределения

X:

x_i	1	2	3	4	5
p_i	0,2	0,3	0,3	0,1	0,1

Найти условную вероятность события $X<5$ при условии, что $X>2$.

3.59. Случайные величины X_1, X_2 независимы и имеют одинаковое распределение

x_i	0	1	2	3
p_i	1/4	1/4	1/4	1/4

а) Найти вероятность события $X_1+X_2> 2$.

б) Найти условную вероятность $P_{X_1=1}[(X_1+X_2)>2]$.

3.60. Распределение дискретной случайной величины X задано формулой $p(X=k) = Ck^2$, где $k = 1, 2, 3, 4, 5$. Найти: а) константу C; б) вероятность события $|X-2|\leq 1$.

3.61. Распределение дискретной случайной величины X определяется формулой

$$P(X=k)=C/2^k, \quad k=0, 1, 2,\ldots .$$

Найти: а) константу C; б) вероятность $P(X \le 3)$.

3.62. Случайная величина X, сосредоточенная на интервале $[-1; 3]$, задана функцией распределения $F(x)=\frac{1}{4}x+\frac{1}{4}$. Найти вероятность попадания случайной величины X в интервал $[0; 2]$. Построить график функции $F(x)$.

3.63. Случайная величина X, сосредоточенная на интервале $[2; 6]$, задана функцией распределения $F(x)=\frac{1}{16}(x^2-4x+4)$. Найти вероятность того, что случайная величина X примет значения: а) меньше 4; б) меньше 6; в) не меньше 3; г) не меньше 6.

3.64. Случайная величина X, сосредоточенная на интервале $(1; 4)$, задана квадратичной функцией распределения $F(x)=ax^2+bx+c$, имеющей максимум при $x=4$. Найти параметры a, b, c и вычислить вероятность попадания случайной величины X в интервал $[2; 3]$.

3.65. Дана функция

$$\varphi(x)=\begin{cases}0 & \text{при } x<0,\\ Cxe^{-x} & \text{при } x\ge 0.\end{cases}$$

При каком значении параметра C эта функция является плотностью распределения некоторой непрерывной случайной величины X? Найти математическое ожидание и дисперсию случайной величины X.

3.66. Случайная величина X задана функцией распределения

$$F(x)=\begin{cases}0 & \text{при } x\le 0,\\ x^2 & \text{при } 0<x\le 1,\\ 1 & \text{при } x>1.\end{cases}$$

Найти: а) плотность вероятности $\varphi(x)$; б) математическое ожидание $M(X)$; в) дисперсию $D(X)$; г) вероятности $P(X=0{,}5)$, $P(X<0{,}5)$, $P(0{,}5\le X\le 1)$; д) построить графики $\varphi(x)$ и $F(x)$ и показать на них математическое ожидание $M(X)$ и вероятности, найденные в п. г).

3.67. По данным примера **3.66** найти: а) моду и медиану случайной величины X; б) квантиль $x_{0,4}$ и 20%-ную точку распределения X.

3.68. По данным примера **3.66** найти коэффициент асимметрии и эксцесс случайной величины X.

3.69. Случайная величина X распределена по закону Коши: $\varphi(x)=\frac{A}{1+x^2}$. Найти: а) коэффициент A; б) функцию распределения $F(x)$; в) вероятность $P(-1\le X\le 1)$. Существуют ли для случайной величины X математическое ожидание и дисперсия?

3.70. Случайная величина X распределена по закону Лапласа: $\varphi(x)=Ae^{-\lambda|x|}$. Найти: а) коэффициент A; б) функцию распределения $F(x)$; в) математическое ожидание $M(X)$ и дисперсию $D(X)$. Построить графики $\varphi(x)$ и $F(x)$.

3.71. Случайная величина X распределена по закону «прямоугольного треугольника» в интервале $(0; c)$ (рис. 3.21). Найти: а) выражение плотности вероятности $\varphi(x)$ и функции распределения $F(x)$; б) математическое ожидание $M(X)$, дисперсию $D(X)$, центральный момент $\mu_3(X)$; в) вероятность $P(c/2\le X\le c)$ и показать ее на данном в условии графике $\varphi(x)$ и построенном графике $F(x)$.

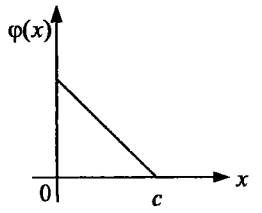


Рис. 3.21

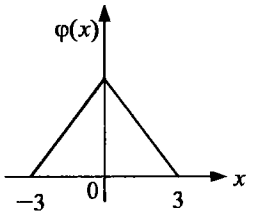


Рис. 3.22

3.72. Случайная величина X распределена по закону Симпсона (равнобедренного треугольника) на отрезке $[-3; 3]$ (рис. 3.22). Найти: а) выражения плотности вероятности $\varphi(x)$ и функции распределения $F(x)$; б) числовые характеристики $M(X)$, $D(X)$, $\mu_3(X)$; в) вероятность $P(-3/2\le X\le 3)$ и показать ее на данном в условии графике $\varphi(x)$ и построенном графике $F(x)$.

Глава 4 Основные законы распределения

В данной главе описаны основные законы распределения дискретных (§ 4.1—4.4) и непрерывных (§ 4.5—4.8) случайных величин, используемых для построения теоретико-вероятностных моделей реальных социально-экономических явлений В § 4.9 рассматриваются распределения случайных величин, используемых в качестве вспомогательного технического средства при решении различных задач статистического анализа.

4.1. Биномиальный закон распределения

Определение. *Дискретная случайная величина X имеет* ***биномиальный закон распределения*** *с параметрами n и p, если она принимает значения* 0, 1, 2,..., *m*,..., *n с вероятностями*

$$P(X=m)=C_n^m p^m q^{n-m}, \qquad (4.1)$$

где $0<p<1$, $q=1-p$.

Как видим, вероятности $P(X=m)$ находятся по формуле Бернулли, полученной выше (гл. 2). Следовательно, *биномиальный закон распределения представляет собой закон распределения числа X=m наступлений события A в n независимых испытаниях, в каждом из которых оно может произойти с одной и той же вероятностью p.*

Ряд распределения биномиального закона имеет вид:

x_i	0	1	2	...	m	...	n
p_i	q^n	$C_n^1 pq^{n-1}$	$C_n^2 p^2 q^{n-2}$	...	$C_n^m p^m q^{n-m}$	...	p^n

Очевидно, что определение биномиального закона корректно, так как основное свойство ряда распределения $\sum_{i=0}^{n} p_i = 1$ выполнено, ибо $\sum_{i=0}^{n} p_i$ есть не что иное, как сумма всех членов разложения бинома Ньютона:

$$q^n + C_n^1 pq^{n-1} + C_n^2 p^2 q^{n-2} + \ldots + C_n^m p^m q^{n-m} + \ldots + p^n = (q+p)^n = 1^n = 1$$

(отсюда и название закона — биномиальный).

На рис. 2.1 приведен многоугольник (полигон) распределения случайной величины X, имеющей биномиальный закон распределения с параметрами $n=5$, $p=0{,}2$, а на переднем форзаце учебника — и при $p=0{,}3$; 0,5; 0,7; 0,8.

Теорема. *Математическое ожидание случайной величины X, распределенной по биномиальному закону,*

$$M(X)=np, \qquad (4.2)$$

а ее дисперсия

$$D(X)=npq. \qquad (4.3)$$

□ Случайную величину X — число m наступлений события A в n независимых испытаниях — можно представить в виде суммы n независимых случайных величин $X_1+X_2+\ldots+X_k+\ldots+X_n$, каждая из которых имеет один и тот же закон распределения, т.е.

$$X=\sum_{k=1}^{n} X_k \text{, где}$$

X_k: ($k=1, 2,\ldots, p$).

x_i	0	1
p_i	q	p

(4.4)

Случайная величина X_k выражает число наступлений события A в k-м (единичном) испытании ($k=1,2,\ldots, n$), т.е. при наступлении события A $X_k=1$ с вероятностью p, при ненаступлении — $X_k=0$ с вероятностью q. Случайную величину X_k называют *альтернативной случайной величиной* (или распределенной *по закону Бернулли, или индикатором события A*).

Найдем числовые характеристики альтернативной случайной величины X_k по формулам **(3.3)** и **(3.11)**.

$$a_k = M(X_k) = \sum_{i=1}^{2} x_i p_i = 0\cdot q + 1\cdot p = p,$$

$$D(X_k) = \sum_{i=1}^{2} (x_i - a_k)^2 p_i = (0-p)^2 q + (1-p)^2 p =$$

$$= p^2 q + q^2 p = pq(p+q) = pq,$$

так как $p+q=1$.

Таким образом, *математическое ожидание альтернативной случайной величины* (**4.4**) *равно вероятности p появления события A в единичном испытании, а ее дисперсия — произведению вероятности p появления события A на вероятность q его непоявления.*

Теперь математическое ожидание и дисперсия рассматриваемой случайной величины X:

$$M(X) = M(X_1+...+X_k+...+X_n) = \underbrace{p+...+p}_{n \text{ раз}} = np,$$

$$D(X) = D(X_1+...+X_k+...+X_n) = \underbrace{pq+...+pq}_{n \text{ раз}} = npq,$$

(при нахождении дисперсии суммы случайных величин учтена их независимость). ■

Следствие. *Математическое ожидание частости* $\frac{m}{n}$ *события в n независимых испытаниях, в каждом из которых оно может наступить с одной и той же вероятностью p, равно p, т.е.*

$$M\left(\frac{m}{n}\right) = p, \qquad (4.5)$$

а *ее дисперсия*

$$D\left(\frac{m}{n}\right) = \frac{pq}{n}. \qquad (4.6)$$

□ Частость события $\frac{m}{n}$ есть $\frac{X}{n}$, т.е. $\frac{m}{n} = \frac{X}{n}$, где X — случайная величина, распределенная по биномиальному закону.

Поэтому

$$M\left(\frac{m}{n}\right) = M\left(\frac{X}{n}\right) = \frac{1}{n}M(X) = \frac{1}{n}\cdot np = p,$$

$$D\left(\frac{m}{n}\right) = D\left(\frac{X}{n}\right) = \frac{1}{n^2}D(X) = \frac{1}{n^2}\cdot npq = \frac{pq}{n}. \blacksquare$$

Замечание. Теперь становится понятным смысл аргументов в функциях $f(x)$ и $\Phi(x)$, содержащихся в локальной и интегральной теоремах Муавра—Лапласа (см. § 2.3). Так, в функции $f(x)$ аргумент $x = \frac{m-np}{\sqrt{npq}}$ есть отклонение числа $X=m$ появления события A в n независимых испытаниях, распределенного по биномиальному закону, от его среднего значения $M(X)=np$, выраженное в стандартных отклонениях $\sigma_x = \sqrt{D(X)} = \sqrt{npq}$. Аргумент $x = \frac{\Delta\sqrt{n}}{\sqrt{pq}} = \frac{\Delta}{\sqrt{pq/n}}$ в функции $\Phi(x)$, рассматриваемой в следствии интегральной теоремы Муавра—Лапласа, есть отклонение Δ частости m/n события A в n независимых испытаниях от его вероятности p в отдельном испытании, выраженное в стандартных отклонениях $\sigma\left(\frac{m}{n}\right) = \sqrt{D\left(\frac{m}{n}\right)} = \sqrt{\frac{pq}{n}}$.

В гл. 2 установлено, что наивероятнейшее число наступлений события A в n повторных независимых испытаниях, в каждом из которых оно может наступить с одной и той же вероятностью p, удовлетворяет неравенству (**2.4**). Это означает, что мода случайной величины, распределенной по биномиальному закону, — число целое — находится из того же неравенства

$$np - q \le Mo(X) \le np + p. \qquad (4.7)$$

Биномиальный закон распределения широко используется в теории и практике статистического контроля качества продукции, при описании функционирования систем массового обслуживания, при моделировании цен активов, в теории стрельбы и в других областях. Так, например, полученный в примере **3.18** закон распределения случайной величины X — числа мальчиков в семье из 4 детей — биномиальный с параметрами $n = 4$, $p = 0{,}515$.

▷ **Пример 4.1.** В магазин поступила обувь с двух фабрик в соотношении 2:3. Куплено 4 пары обуви. Найти закон распределения числа купленных пар обуви, изготовленной первой фабрикой. Найти математическое ожидание и среднее квадратическое отклонение этой случайной величины.

Р е ш е н и е. Вероятность того, что случайно выбранная пара обуви изготовлена первой фабрикой, равна p=2/(2+3)=0,4. Случайная величина X — число пар обуви среди четырех, изготовленных первой фабрикой, имеет биномиальный закон распределения с параметрами $n = 4$, $p = 0{,}4$. Ряд распределения X имеет вид:

x_i	0	1	2	3	4
p_i	0,1296	0,3456	0,3456	0,1536	0,0256

(Значения $p_i=P(X=m)$, $(m=0,1,2,3,4)$ вычислены по формуле **(4.1)**: $P(X=m)=C_4^m\cdot 0{,}4^m\cdot 0{,}6^{4-m}$.

Найдем математическое ожидание и дисперсию случайной величины X по формулам **(4.2)** и **(4.3)**:

$$M(X)=np=4\cdot 0{,}4=1{,}6,\ D(X)=npq=4\cdot 0{,}4\cdot 0{,}6=0{,}96.$$

Замечание. Нетрудно заметить, что полученное распределение двумодальное (имеющее две моды): $Mo(X)_1=1$ и $Mo(X)_2=2$, так как эти значения имеют наибольшие (и равные между собой) вероятности. Моду $Mo(X)$ — число целое — можно найти из неравенства **(4.7)**: $4\cdot 0{,}4-0{,}6\le Mo(X)\le 4\cdot 0{,}4+0{,}4$ или $1\le Mo(X)\le 2$, т.е. $Mo(X)_1=1$ и $Mo(X)_2=2$. ▶

▷**Пример 4.2.** По данным примера **4.1** найти математическое ожидание и дисперсию частости (доли) пар обуви, изготовленных первой фабрикой, среди 4 купленных.

Решение. Имеем $n=4$, $p=0{,}4$. По формулам **(4.5)**, **(4,6)**:

$$M\left(\frac{m}{n}\right)=0{,}4,\quad D\left(\frac{m}{n}\right)=\frac{0{,}4\cdot 0{,}6}{4}=0{,}06.\ ▶$$

4.2. Закон распределения Пуассона

Определение. *Дискретная случайная величина X имеет **закон распределения Пуассона** с параметром $\lambda>0$, если она принимает значения 0, 1, 2,..., m,... (бесконечное, но счетное множество значений) с вероятностями*

$$P(X=m)=\frac{\lambda^m e^{-\lambda}}{m!}=P_m(\lambda),\qquad (4.8)$$

Ряд распределения закона Пуассона имеет вид:

x_i	0	1	2	...	m	...
p_i	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{\lambda^2 e^{-\lambda}}{2!}$	...	$\frac{\lambda^m e^{-\lambda}}{m!}$	...

Очевидно, что определение закона Пуассона корректно, так как основное свойство ряда распределения $\sum_{i=1}^{\infty}p_i=1$ выполнено, ибо сумма ряда

$$\sum_{i=1}^{\infty}p_i=e^{-\lambda}+\lambda e^{-\lambda}+\frac{\lambda^2e^{-\lambda}}{2!}+\ldots+\frac{\lambda^m e^{-\lambda}}{m!}+\ldots=$$
$$=e^{-\lambda}\left(1+\lambda+\frac{\lambda^2}{2!}+\ldots+\frac{\lambda^m}{m!}+\ldots\right)=e^{-\lambda}\cdot e^{\lambda}=1$$

(учтено, что в скобках записано разложение в ряд функции e^x при $x=\lambda$).

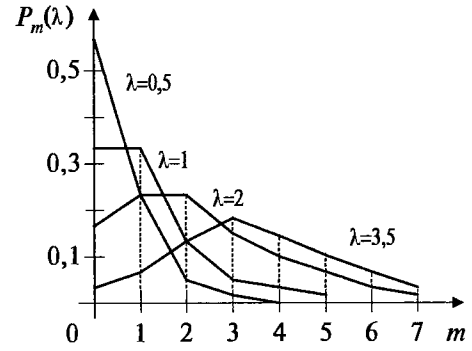


Рис. 4.1

На рис. 4.1 показан многоугольник (полигон) распределения случайной величины, распределенной по закону Пуассона $P(X=m)=P_m(\lambda)$ с параметрами $\lambda=0{,}5$, $\lambda=1$, $\lambda=2$, $\lambda=3{,}5$.

Теорема. *Математическое ожидание и дисперсия случайной величины, распределенной по закону Пуассона, совпадают и равны параметру λ этого закона*, т.е.

$$M(X)=\lambda,\qquad (4.9)$$

$$D(X)=\lambda.\qquad (4.10)$$

□ Найдем математическое ожидание случайной величины X:

$$a=M(X)=\sum_{i=1}^{\infty}x_ip_i=\sum_{m=0}^{\infty}m\frac{\lambda^m e^{-\lambda}}{m!}=\sum_{m=1}^{\infty}\frac{\lambda^m e^{-\lambda}}{(m-1)!}=$$
$$=\lambda e^{-\lambda}\sum_{m=1}^{\infty}\frac{\lambda^{m-1}}{(m-1)!}=\lambda e^{-\lambda}\left(1+\lambda+\frac{\lambda^2}{2!}+\ldots\right)=\lambda e^{-\lambda}e^{\lambda}=\lambda.$$

Дисперсию случайной величины X найдем по формуле **(3.16)**, т.е. $D(X)=M(X^2)-a^2$. Вначале получим формулу для

$$M(X^2)=\sum_{i=1}^{\infty}x_i^2p_i=\sum_{m=0}^{\infty}m^2\frac{\lambda^m e^{-\lambda}}{m!}=\sum_{m=1}^{\infty}m\frac{\lambda^m e^{-\lambda}}{(m-1)!}=$$

$$=e^{-\lambda}\sum_{m=1}^{\infty}\frac{[(m-1)+1]\lambda^m}{(m-1)!}=\lambda^2e^{-\lambda}\sum_{m=2}^{\infty}\frac{\lambda^{m-2}}{(m-2)!}+\lambda e^{-\lambda}\sum_{m=1}^{\infty}\frac{\lambda^{m-1}}{(m-1)!}=$$

$$=\lambda^2e^{-\lambda}\left(1+\lambda+\frac{\lambda^2}{2!}+\ldots\right)+\lambda e^{-\lambda}\left(1+\lambda+\frac{\lambda^2}{2!}+\ldots\right)=$$

$$=\lambda^2e^{-\lambda}e^{\lambda}+\lambda e^{-\lambda}e^{\lambda}=\lambda^2+\lambda.$$

Теперь $D(X)=(\lambda^2+\lambda)-\lambda^2=\lambda$. ■

При достаточно больших n (вообще при $n\to\infty$) и малых значениях p ($p\to 0$) при условии, что произведение np — постоянная величина ($np\to\lambda=\text{const}$), закон распределения Пуассона является хорошим приближением биномиального закона, так как в этом случае функция вероятностей Пуассона **(4.8)** хорошо аппроксимирует функцию вероятностей **(4.1)**, определяемую по формуле Бернулли (см. § 2.2). Иначе, при $p\to 0$, $n\to\infty$, $np\to\lambda=\text{const}$ *закон распределения Пуассона является предельным случаем биномиального закона.* Так как при этом вероятность p события A в каждом испытании мала, то закон распределения Пуассона называют часто *законом редких явлений.*

Наряду с «предельным» случаем биномиального распределения закон Пуассона может возникнуть и в ряде других ситуаций. Так, в гл. 7 показано, что для простейшего потока событий число событий, попадающих на произвольный отрезок времени, есть случайная величина, имеющая пуассоновское распределение.

По закону Пуассона распределены, например, число рождения четверней, число сбоев на автоматической линии, число отказов сложной системы в «нормальном режиме», число «требований на обслуживание», поступивших в единицу времени в системах массового обслуживания, и др.

Отметим еще, что если случайная величина представляет собой сумму двух независимых случайных величин, распределенных каждая по закону Пуассона, то она также распределена по закону Пуассона.

▷ **Пример 4.3.** Доказать, что сумма двух независимых случайных величин, распределенных по закону Пуассона с параметрами λ_1 и λ_2, также распределена по закону Пуассона с параметром $\lambda=\lambda_1+\lambda_2$.

Р е ш е н и е. Пусть случайные величины $X=m$ и $Y=n$ имеют законы распределения Пуассона соответственно с параметрами λ_1 и λ_2. В силу независимости случайных величин X и Y их сумма $Z=X+Y$ принимает значение $Z=s$ с вероятностью

$$P(Z=s)=P(X=m)\cdot P(Y=n)=$$

$$=\sum_{m+n=s}\frac{\lambda_1^m e^{-\lambda_1}}{m!}\cdot\frac{\lambda_2^n e^{-\lambda_2}}{n!}=e^{-(\lambda_1+\lambda_2)}\sum_{m+n=s}\frac{\lambda_1^m\lambda_2^n}{m!\,n!}=$$

$$=e^{-(\lambda_1+\lambda_2)}\sum_{n=0}^{s}\frac{\lambda_1^{s-n}\cdot\lambda_2^n}{(s-n)!\,n!}=\frac{e^{-(\lambda_1+\lambda_2)}}{s!}\sum_{n=0}^{s}\frac{s!}{(s-n)!\,n!}\lambda_1^{s-n}\lambda_2^n.$$

Полагая, что $\lambda_1+\lambda_2=\lambda$, и учитывая, что $\sum_{n=0}^{s}\frac{s!}{(s-n)!\,n!}\lambda_1^{s-n}\lambda_2^n=$

$=\sum_{n=0}^{s}C_s^n\lambda_1^{s-n}\lambda_2^n=(\lambda_1+\lambda_2)^s=\lambda^s$, получим $P(Z=s)=\frac{e^{-\lambda}\lambda^s}{s!}$, т.е. случайная величина $Z=X+Y$ распределена по закону Пуассона с параметром $\lambda=\lambda_1+\lambda_2$. ▶

4.3. Геометрическое распределение

О п р е д е л е н и е. *Дискретная случайная величина $X=m$ имеет **геометрическое распределение** с параметром p, если она принимает значения* 1,2,..., m... (бесконечное, но счетное множество значений) с вероятностями

$$P(X=m)=pq^{m-1}, \qquad \textbf{(4.11)}$$

где $0<p<1$, $q=1-p$.

Ряд геометрического распределения случайной величины имеет вид:

x_i	1	2	3	...	m	...
p_i	p	pq	pq^2	...	pq^{m-1}	...

Нетрудно видеть, что вероятности p_i образуют геометрическую прогрессию с первым членом p и знаменателем q (отсюда название «геометрическое распределение»).

Определение геометрического распределения корректно, так как сумма ряда

$$\sum_{i=1}^{\infty} p_i = p + pq + \ldots + pq^{m-1} + \ldots = p\left(1 + q + \ldots + q^{m-1} + \ldots\right) = p\frac{1}{1-q} = \frac{p}{p} = 1$$

(так как $\frac{1}{1-q} = \frac{1}{p}$ есть сумма геометрического ряда $\sum_{m=1}^{\infty} q^{m-1}$ при $|q| < 1$).

Случайная величина X=m, имеющая геометрическое распределение, представляет собой число m испытаний, проведенных по схеме Бернулли, с вероятностью p наступления события в каждом испытании до первого положительного исхода.

Так, например, число вызовов радистом корреспондента до тех пор, пока вызов не будет принят, рассматриваемое в примере **3.19***б*, есть случайная величина, имеющая геометрическое распределение с параметром $p = 0{,}4$.

Теорема. *Математическое ожидание случайной величины X, имеющей геометрическое распределение с параметром p,*

$$M(X) = \frac{1}{p}, \qquad \textbf{(4.12)}$$

а ее дисперсия[1]

$$D(X) = \frac{q}{p^2}, \qquad \textbf{(4.13)}$$

где $q = 1 - p$.

▷ **Пример 4.4.** Проводится проверка большой партии деталей до обнаружения бракованной (без ограничения числа проверенных деталей). Составить закон распределения числа проверенных деталей. Найти его математическое ожидание и дисперсию, если известно, что вероятность брака для каждой детали равна 0,1.

Р е ш е н и е. Случайная величина X — число проверенных деталей до обнаружения бракованной — имеет геометрическое распределение **(4.11)** с параметром $p = 0{,}1$. Поэтому ряд распределения имеет вид

X=m:	x_i	1	2	3	4	...	m	...
	p_i	0,1	0,09	0,081	0,0729	...	$0{,}9^m \cdot 0{,}1$	...

По формулам **(4.12)** и **(4.13)**

$$M(X) = \frac{1}{p} = \frac{1}{0{,}1} = 10,\ D(X) = \frac{q}{p^2} = \frac{0{,}9}{0{,}1^2} = 90$$ ▶

4.4. Гипергеометрическое распределение

О п р е д е л е н и е. *Дискретная случайная величина X имеет **гипергеометрическое распределение** с параметрами n, M, N, если она принимает значения*[1] 0, 1, 2, m,..., min (n, M) *с вероятностями*

$$P(X = m) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}, \qquad \textbf{(4.14)}$$

где $M \le N$, $n \le N$; n, M, N — *натуральные числа.*

Гипергеометрическое распределение имеет случайная величина X=m — число объектов, обладающих заданным свойством, среди n объектов, случайно извлеченных (без возврата) из совокупности N объектов, M из которых обладают этим свойством.

Так, распределение случайной величины X — числа неточных приборов среди взятых наудачу четырех, полученное в примере **3.20**, есть гипергеометрическое распределение с параметрами n=4, M=3, N=10.

Теорема. *Математическое ожидание случайной величины X, имеющей гипергеометрическое распределение с параметрами n, M, N, есть*

$$M(X) = n\frac{M}{N}, \qquad \textbf{(4.15)}$$

[1] Доказательство теоремы, связанное с суммированием членов бесконечного ряда, здесь не приводим. Это доказательство аналогично приведенному для частного случая в решении примера **3.19***б*.

[1] Точнее, возможные значения m заключены в границах от max (0, $n + M - N$) до min (n, M), при которых существуют C_n^m, C_{N-M}^{n-m}

а *ее дисперсия*

$$D(X)=n\frac{M}{N-1}\left(1-\frac{M}{N}\right)\left(1-\frac{n}{N}\right). \qquad \textbf{(4.16)}$$

Случайную величину $X=m$, распределенную по биномиальному закону **(4.1)**, можно интерпретировать как число m объектов, обладающих данным свойством, из общего числа n объектов, случайно извлеченных из некоторой воображаемой б е с к о н е ч н о й совокупности, доля p объектов которой обладает этим свойством. Поэтому *гипергеометрическое распределение можно рассматривать как модификацию биномиального распределения для случая к о н е ч н о й совокупности, состоящей из N объектов, M из которых обладают этим свойством.*

Можно показать, что при $N\to\infty$ функция вероятностей **(4.14)** гипергеометрического распределения стремится к соответствующей функции **(4.1)** биномиального закона.

Гипергеометрическое распределение широко используется в практике статистического приемочного контроля качества промышленной продукции, в задачах, связанных с организацией выборочных обследований, и других областях.

▷ **Пример 4.5.** В лотерее «Спортлото 6 из 45» денежные призы получают участники, угадавшие 3, 4, 5 и 6 видов спорта из отобранных случайно 6 видов из 45 (размер приза увеличивается с увеличением числа угаданных видов спорта). Найти закон распределения случайной величины X — числа угаданных видов спорта среди случайно отобранных шести. Какова вероятность получения денежного приза? Найти математическое ожидание и дисперсию случайной величины X.

Р е ш е н и е. Очевидно (см. гл. 1, пример **1.14**), что число угаданных видов спорта в лотерее «6 из 45» есть случайная величина, имеющая гипергеометрическое распределение с параметрами $n = 6$, $M = 6$, $N = 45$. Ряд ее распределения, рассчитанный по формуле **(4.14)**, имеет вид:

X:	x_i	0	1	2	3	4	5	6
	p_i	0,40056	0,42413	0,15147	0,02244	0,00137	0,00003	0,0000001

Вероятность получения денежного приза

$$P(3\le X\le 6)=\sum_{i=3}^{6}P(X=i)=$$

$$=0{,}02244+0{,}00137+0{,}00003+0{,}0000001=0{,}02384\approx 0{,}024.$$

По формулам **(4.15)** и **(4.16)**

$$M(X)=6\cdot\frac{6}{45}=0{,}8;\quad D(X)=6\cdot\frac{39}{44}\left(1-\frac{39}{45}\right)\left(1-\frac{6}{45}\right)=0{,}6145.$$

Таким образом, среднее число угаданных видов спорта из 6 всего 0,8, а вероятность выигрыша только 0,024 ▶

4.5. Равномерный закон распределения

О п р е д е л е н и е. *Непрерывная случайная величина X имеет **равномерный закон** распределения на отрезке $[a, b]$, если ее плотность вероятности $\varphi(x)$ постоянна на этом отрезке и равна нулю вне его*, т.е.

$$\varphi(x)=\begin{cases}\dfrac{1}{b-a} & \text{при } a\le x\le b,\\ 0 & \text{при } x<a,\ x>b.\end{cases} \qquad \textbf{(4.17)}$$

Кривая распределения $\varphi(x)$ и график функции распределения $F(x)$ случайной величины X приведены на рис. 4.2 *а, б*.

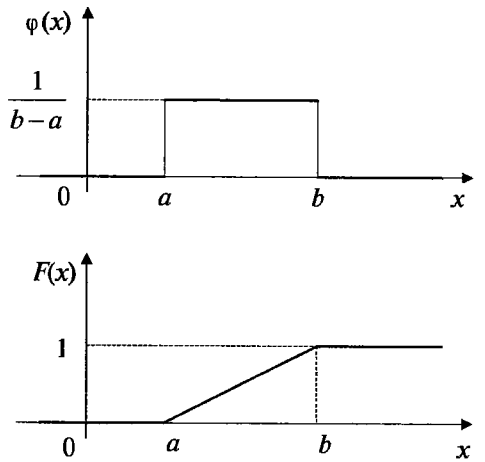


Рис. 4.2

Теорема. *Функция распределения случайной величины X, распределенной по равномерному закону, есть*

$$F(x)=\begin{cases}0 & \text{при } x\le a,\\ (x-a)\,(b-a) & \text{при } a<x\le b,\\ 1 & \text{при } x>b,\end{cases} \qquad \textbf{(4.18)}$$

ее математическое ожидание

$$M(X)=\frac{a+b}{2}, \qquad (4.19)$$

а *дисперсия*

$$D(X)=\frac{(b-a)^2}{12}. \qquad (4.20)$$

□ При $x \le a$ функция распределения $F(x)=0$.

При $a < x \le b$ по формуле **(3.23)**

$$F(x)=\int_a^x \frac{dx}{b-a}=\frac{x}{b-a}\Big|_a^x=\frac{x-a}{b-a}.$$

При $x>b$ очевидно, что

$$F(x)=\int_a^b \frac{dx}{b-a}=\frac{b-a}{b-a}=1,$$

т.е. формула **(4.18)** доказана.

Математическое ожидание случайной величины X с учетом его механической интерпретации как центра массы равно абсциссе середины отрезка, т.е. $M(X)=\frac{a+b}{2}$.

Тот же результат получается по формуле **(3.25)**:

$$M(X)=\int_{-\infty}^{+\infty} x\varphi(x)dx=\int_a^b \frac{x\,dx}{b-a}=\frac{1}{b-a}\left(\frac{x^2}{2}\Big|_a^b\right)=\frac{b^2-a^2}{2(b-a)}=\frac{a+b}{2}.$$

По формуле **(3.26)**:

$$D(X)=\int_{-\infty}^{+\infty}\left[X-M(X)\right]^2\varphi(x)dx=$$

$$=\int_a^b\left(x-\frac{a+b}{2}\right)^2\frac{dx}{b-a}=\frac{1}{3(b-a)}\left(x-\frac{a+b}{2}\right)^3\Big|_a^b=$$

$$=\frac{1}{3(b-a)}\left[\frac{(b-a)^3}{8}-\frac{(a-b)^3}{8}\right]=\frac{(b-a)^2}{12}. \blacksquare$$

Равномерный закон распределения используется при анализе ошибок округления при проведении числовых расчетов (например, ошибка округления числа до целого распределена равномерно на отрезке $[-0{,}5;\ +0{,}5]$), в ряде задач массового обслуживания, при статистическом моделировании наблюдений, подчиненных заданному распределению. Так, случайная величина X, распределенная по равномерному закону на отрезке $[0;1]$, называемая

«случайным числом от 0 до 1», служит исходным материалом для получения случайных величин с любым законом распределения.

▷ **Пример 4.6.** Поезда метрополитена идут регулярно с интервалом 2 мин. Пассажир выходит на платформу в случайный момент времени. Какова вероятность того, что ждать пассажиру придется не больше полминуты. Найти математическое ожидание и среднее квадратическое отклонение случайной величины X — времени ожидания поезда.

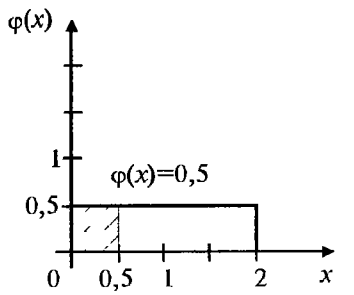


Рис. 4.3

Р е ш е н и е. Случайная величина X — время ожидания поезда на временном (в минутах) отрезке $[0;2]$ имеет равномерный закон распределения $\varphi(x)=\frac{1}{2}$.

Поэтому вероятность того, что пассажиру придется ждать не более полминуты, равна 1/4 от равной единице площади прямоугольника (рис. 4.3), т.е.

$$P(X\le 0{,}5)=\int_0^{0{,}5}\frac{1}{2}dx=\frac{1}{2}x\Big|_0^{0{,}5}=\frac{1}{4}.$$

По формулам **(4.19)** и **(4.20)**

$$M(X)=\frac{0+2}{2}=1 \text{ мин}, \quad D(X)=\frac{(2-0)^2}{12}=\frac{1}{3},$$

$$\sigma_x=\sqrt{D(X)}=\sqrt{\frac{1}{3}}=\frac{\sqrt{3}}{3}\approx 0{,}58 \text{ мин}. \blacktriangleright$$

4.6. Показательный (экспоненциальный) закон распределения

О п р е д е л е н и е. *Непрерывная случайная величина X имеет* ***показательный (экспоненциальный)*** *закон распределения с параметром $\lambda > 0$, если ее плотность вероятности имеет вид:*

$$\varphi(x)=\begin{cases}\lambda e^{-\lambda x} & \text{при } x\ge 0,\\ 0 & \text{при } x<0.\end{cases} \qquad (4.21)$$

Кривая распределения $\varphi(x)$ и график функции распределения $F(x)$ случайной величины X приведены на рис. 4.4 *а, б.*

Теорема. *Функция распределения случайной величины X, распределенной по показательному (экспоненциальному) закону, есть*

$$F(x)=\begin{cases}0 & \text{при } x<0,\\ 1-e^{-\lambda x} & \text{при } x\ge 0,\end{cases} \qquad \textbf{(4.22)}$$

ее *математическое ожидание*

$$M(X)=\frac{1}{\lambda}, \qquad \textbf{(4.23)}$$

а *дисперсия*

$$D(X)=\frac{1}{\lambda^2} \qquad \textbf{(4.24)}$$

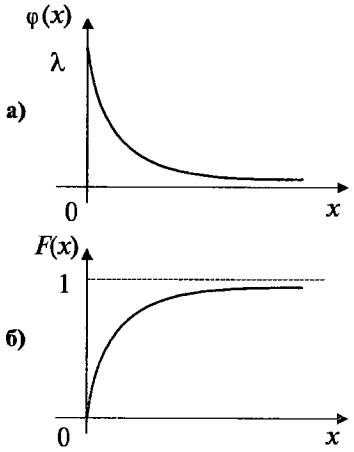


Рис. 4.4

□ При $x<0$ функция распределения $F(x) = 0$. При $x \ge a$ по формуле **(3.23)**

$$F(x)=\int_0^x \lambda e^{-\lambda x}dx=-e^{-\lambda x}\Big|_0^x=1-e^{-\lambda x},$$

т.е. формула **(4.22)** доказана.

Найдем математическое ожидание случайной величины X, используя при вычислении метод интегрирования по частям:

$$a=M(X)=\int_{-\infty}^{+\infty} x\,\varphi(x)dx=$$

$$=\lim_{b\to\infty}\int_0^b x\,\lambda e^{-\lambda x}dx=\lim_{b\to+\infty}\left(-\int_0^b x\,de^{-\lambda x}\right)=$$

$$=\lim_{b\to+\infty}\left(-xe^{-\lambda x}\Big|_0^b+\int_0^b e^{-\lambda x}dx\right)=\lim_{b\to+\infty}\left(-be^{-\lambda b}-\frac{1}{\lambda}e^{-\lambda x}\Big|_0^b\right)=$$

$$=0-\frac{1}{\lambda}\lim_{b\to+\infty}\left(e^{-\lambda b}-1\right)=\frac{1}{\lambda}.$$

Для нахождения дисперсии $D(X)$ вначале найдем

$$M(X^2)=\int_{-\infty}^{+\infty}x^2\varphi(x)dx=\lim_{b\to+\infty}\int_0^b x^2\lambda e^{-\lambda x}dx=$$

$$=\lim_{b\to+\infty}\left(-\int_0^b x^2de^{-\lambda x}\right)=\lim_{b\to+\infty}\left(-x^2e^{-\lambda x}\Big|_0^b+\int_0^b 2x\,e^{-\lambda x}dx\right)=$$

$$=\lim_{b\to+\infty}\left(-b^2e^{-\lambda b}\right)+\frac{2}{\lambda}\lim_{b\to+\infty}\int_0^b x\lambda^{-\lambda x}dx=0+\frac{2}{\lambda}\cdot\frac{1}{\lambda}=\frac{2}{\lambda^2},$$

с учетом того, что во втором слагаемом несобственный интеграл есть $M(X)=\frac{1}{\lambda}$. Теперь

$$D(X)=M(X^2)-a^2=\frac{2}{\lambda^2}-\frac{1}{\lambda^2}=\frac{1}{\lambda^2}. \blacksquare$$

Из доказанной теоремы следует, что для *случайной величины, распределенной по показательному закону, математическое ожидание равно среднему квадратическому отклонению,* т.е.

$$M(X)=\sigma_x=1/\lambda.$$

Показательный закон распределения играет большую роль в теории массового обслуживания и теории надежности. Так, например, интервал времени T между двумя соседними событиями в простейшем потоке имеет показательное распределение с параметром λ — интенсивностью потока (см. § 7.3).

Показательный закон распределения (и только он) обладает важным свойством, рассматриваемым ниже.

▷ **Пример 4.7.** Доказать, что если промежуток времени T, распределенный по показательному закону, уже длился некоторое время τ, то это никак не влияет на закон распределения оставшейся части $T_1 = T-\tau$ промежутка, т.е. закон распределения T_1 остается таким же, как и всего промежутка T.

Р е ш е н и е. Пусть функция распределения промежутка T определяется по формуле **(4.22)**, т.е. $F(t) = 1-e^{-\lambda t}$, а функция распределения оставшейся части $T_1 = T-\tau$ при условии, что событие $T > \tau$ произошло, есть условная вероятность события $T_1 < t$ относительно события $T > \tau$, т.е. $F_1(t) = P_{T>\tau}(T_1 < t)$.

Так как условная вероятность любого события B относительно события A $P_A(B)=P(AB)/P(A)$, то, полагая $A=(T>\tau)$, $B=(T_1<t)$, получим

$$F_1(t)=P_{T>\tau}(T_1<t)=\frac{P[(T>\tau)(T_1<t)]}{P(T>\tau)}. \qquad (4.25)$$

Произведение событий $(T>\tau)$ и $T_1=T-\tau<t$ равносильно событию $\tau<T<t+\tau$, вероятность которого

$$P(\tau<T<t+\tau)=F(t+\tau)-F(\tau).$$

Так как $P(T>\tau)=1-P(T\le\tau)=1-F(\tau)$, то выражение **(4.25)** можно представить в виде:

$$F_1(t)=\frac{F(t+\tau)-F(\tau)}{1-F(\tau)}.$$

Учитывая **(4.22)**, получим

$$F_1(t)=\frac{e^{-\lambda\tau}-e^{-\lambda(t+\tau)}}{e^{-\lambda\tau}}=1-e^{-\lambda t}=F(t). \blacktriangleright$$

Доказанное в примере **4.7** свойство показательного распределения широко используется в марковских случайных процессах (см. гл. 7).

▷ **Пример 4.8.** Установлено, что время ремонта телевизоров есть случайная величина X, распределенная по показательному закону. Определить вероятность того, что на ремонт телевизора потребуется не менее 20 дней, если среднее время ремонта телевизоров составляет 15 дней. Найти плотность вероятности, функцию распределения и среднее квадратическое отклонение случайной величины X.

Р е ш е н и е. По условию математическое ожидание $M(X)=\frac{1}{\lambda}=15$, откуда параметр $\lambda=1/15$ и по формулам **(4.21)** и **(4.22)** плотность вероятности и функция распределения имеют вид:

$$\varphi(x)=\frac{1}{15}e^{-\frac{1}{15}x};\quad F(x)=1-e^{-\frac{1}{15}x}\ (x\ge 0).$$

Искомую вероятность $P(X\ge 20)$ можно было найти по формуле **(3.22)**, интегрируя плотность вероятности, т.е.

$$P(X\ge 20)=P(20\le X<+\infty)=\int_{20}^{+\infty}\frac{1}{15}e^{-\frac{1}{15}x}dx,$$

но проще это сделать, используя функцию распределения:

$$P(X\ge 20)=1-P(X<20)=1-F(20)=1-(1-e^{-\frac{20}{15}})=e^{-\frac{20}{15}}=0{,}264.$$

Осталось найти среднее квадратическое отклонение $\sigma_x=M(X)=15$ дней. ▶

4.7. Нормальный закон распределения

Нормальный закон распределения наиболее часто встречается на практике. Главная особенность, выделяющая его среди других законов, состоит в том, что он является п р е д е л ь н ы м законом, к которому приближаются другие законы распределения при весьма часто встречающихся типичных условиях (см. гл. 6).

О п р е д е л е н и е. *Непрерывная случайная величина X имеет **нормальный закон распределения (закон Гаусса)** с параметрами a и σ^2, если ее плотность вероятности имеет вид:*

$$\varphi_N(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}. \qquad (4.26)$$

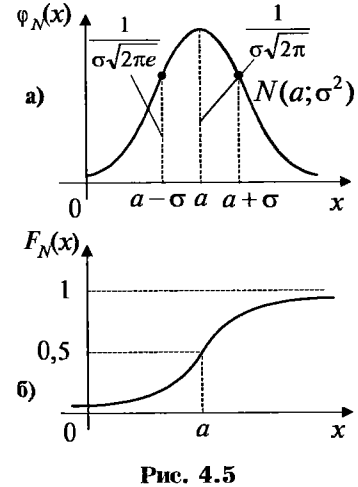


Рис. 4.5

Термин «нормальный» не совсем удачный. Многие признаки подчиняются нормальному закону, например, рост человека, дальность полета снаряда и т.п. Но если какой-либо признак подчиняется другому, отличному от нормального, закону распределения, то это вовсе не говорит о «ненормальности» явления, связанного с этим признаком.

Кривую нормального закона распределения называют *нормальной* или *гауссовой кривой*. На рис. 4.5 *а*, *б* приведены нормальная кривая $\varphi_N(x)$ с параметрами a и σ^2, т.е.

$N(a;\sigma^2)$, и график функции распределения случайной величины X, имеющей нормальный закон. Обратим внимание на то, что нормальная кривая симметрична относительно прямой $x=a$, имеет максимум в точке $x=a$, равный $1/(\sigma\sqrt{2\pi})$, т.е. $f_{\max}(a)=\dfrac{1}{\sigma\sqrt{2\pi}}\approx\dfrac{0{,}3989}{\sigma}$, и две точки перегиба $x=a\pm\sigma$ с ординатой $f_{\text{пер}}(a\pm\sigma)=\dfrac{1}{\sigma\sqrt{2\pi e}}\approx\dfrac{0{,}2420}{\sigma}$.

Можно заметить, что в выражении плотности нормального закона параметры обозначены буквами a и σ^2, которыми мы обозначаем математическое ожидание $M(X)$ и дисперсию $D(X)$. Такое совпадение неслучайно. Рассмотрим теорему, устанавливающую теоретико-вероятностный смысл параметров нормального закона.

Теорема. *Математическое ожидание случайной величины X, распределенной по нормальному закону, равно параметру a этого закона*, т.е.

$$M(X)=a, \qquad \textbf{(4.27)}$$

а *ее дисперсия — параметру* σ^2, т.е.

$$D(X)=\sigma^2. \qquad \textbf{(4.28)}$$

□ Математическое ожидание случайной величины X:

$$M(X)=\int_{-\infty}^{+\infty}x\,\varphi_N(x)dx=\int_{-\infty}^{+\infty}x\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}dx.$$

Произведем замену переменной, положив $t=\dfrac{x-a}{\sigma\sqrt{2}}$.

Тогда $x=a+\sigma\sqrt{2}t$ и $dx=\sigma\sqrt{2}dt$, пределы интегрирования не меняются и, следовательно,

$$M(X)=\int_{-\infty}^{+\infty}\frac{1}{\sigma\sqrt{2\pi}}(a+\sigma\sqrt{2}t)e^{-t^2}\sigma\sqrt{2}\,dt=$$

$$=\frac{\sigma\sqrt{2}}{\sqrt{\pi}}\int_{-\infty}^{+\infty}te^{-t^2}dt+\frac{a}{\sqrt{\pi}}\int_{-\infty}^{+\infty}e^{-t^2}dt=0+\frac{a}{\sqrt{\pi}}\cdot\sqrt{\pi}=a$$

(первый интеграл равен нулю как интеграл от нечетной функции по симметричному относительно начала координат промежутку, а второй интеграл $\int_{-\infty}^{+\infty}e^{-t^2}dt=\sqrt{\pi}$ — интеграл Эйлера—Пуассона).

Дисперсия случайной величины X:

$$D(X)=\int_{-\infty}^{+\infty}(x-a)^2\varphi_N(x)dx=\int_{-\infty}^{+\infty}(x-a)^2\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}dx.$$

Сделаем ту же замену переменной $x=a+\sigma\sqrt{2}t$, как и при вычислении предыдущего интеграла. Тогда

$$D(X)=\int_{-\infty}^{+\infty}\sigma^2 2t^2\frac{1}{\sigma\sqrt{2\pi}}e^{-t^2}\sigma\sqrt{2}\,dt=\frac{2\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{+\infty}t^2e^{-t^2}dt=-\frac{\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{+\infty}t\,de^{-t^2}.$$

Применяя метод интегрирования по частям, получим

$$D(X)=-\frac{\sigma^2}{\sqrt{\pi}}te^{-t^2}\Big|_{-\infty}^{+\infty}+\frac{\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{+\infty}e^{-t^2}dt=0+\frac{\sigma^2}{\sqrt{\pi}}\cdot\sqrt{\pi}=\sigma^2.\ \blacktriangleright$$

Выясним, как будет меняться нормальная кривая при изменении параметров a и σ^2 (или σ). Если $\sigma=\text{const}$, и меняется параметр a ($a_1<a_2<a_3$), т.е. центр симметрии распределения, то нормальная кривая будет смещаться вдоль оси абсцисс, не меняя формы (рис. 4.6).

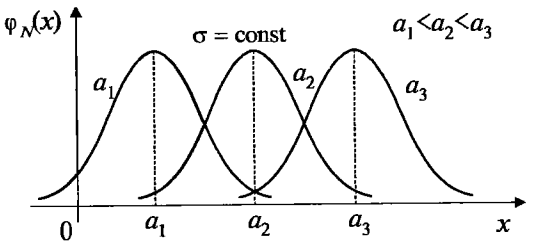


Рис. 4.6

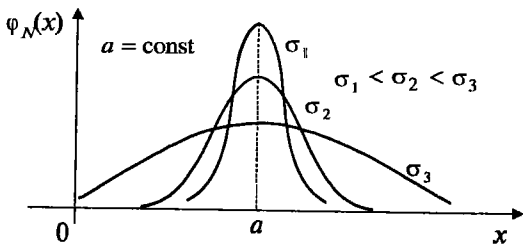


Рис. 4.7

Если a=const и меняется параметр σ^2 (или σ), то меняется ордината максимума кривой $f_{\max}(a) = \dfrac{1}{\sigma\sqrt{2\pi}}$. При увеличении σ ордината максимума кривой уменьшается, но так как площадь под любой кривой распределения должна оставаться равной единице, то кривая становится более плоской, растягиваясь вдоль оси абсцисс; при уменьшении σ, напротив, нормальная кривая вытягивается вверх, одновременно сжимаясь с боков. На рис. 4.7 показаны нормальные кривые с параметрами σ_1, σ_2 и σ_3, где $\sigma_1 < \sigma_2 < \sigma_3$. Таким образом, параметр a (он же математическое ожидание) характеризует положение центра, а параметр σ^2 (он же дисперсия) — форму нормальной кривой.

Нормальный закон распределения случайной величины с параметрами a=0, σ^2=1, т.е. $N(0;1)$, называется *стандартным* или *нормированным*, а соответствующая нормальная кривая — *стандартной* или *нормированной*.

Сложность непосредственного нахождения функции распределения случайной величины, распределенной по нормальному закону, по формуле **(3.23)** и вероятности ее попадания на некоторый промежуток по формуле **(3.22)** связана с тем, что интеграл от функции **(4.26)** является «неберущимся» в элементарных функциях. Поэтому их выражают через функцию

$$\Phi(x) = \frac{2}{\sqrt{2\pi}}\int_0^x e^{-t^2/2}dt \qquad \textbf{(4.29)}$$

— *функцию* (*интеграл вероятностей*) *Лапласа*, для которой составлены таблицы. Напомним, что функция Лапласа уже встречалась нам при рассмотрении интегральной теоремы Муавра—Лапласа (см. § 2.3). Там же были рассмотрены ее свойства. Геометрически функция Лапласа представляет собой площадь под стандартной нормальной кривой на отрезке $[-x; x]$ (рис. 4.8).

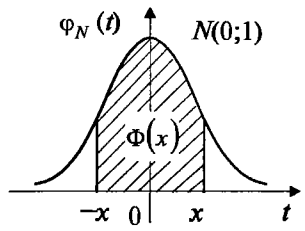


Рис. 4.8

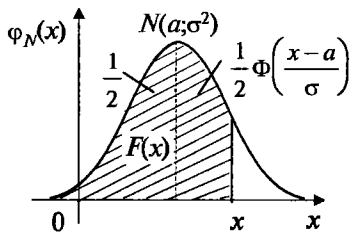


Рис. 4.9

Теорема. *Функция распределения случайной величины X, распределенной по нормальному закону, выражается через функцию Лапласа* $\Phi(x)$ по формуле:

$$F_N(x) = \frac{1}{2} + \frac{1}{2}\Phi\left(\frac{x-a}{\sigma}\right). \qquad \textbf{(4.30)}$$

□ По формуле **(3.23)** функция распределения:

$$F_N(x) = \int_{-\infty}^{x}\varphi_N(x)dx = \int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}dx. \qquad \textbf{(4.31)}$$

Сделаем замену переменной, полагая $t = \dfrac{x-a}{\sigma}$, $x = a + t\sigma$, $dx = \sigma dt$; при $x \to -\infty$ $t \to -\infty$, поэтому

$$F_N(x) = \int_{-\infty}^{\frac{x-a}{\sigma}}\frac{1}{\sigma\sqrt{2\pi}}e^{-t^2/2}\sigma dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{x-a}{\sigma}}e^{-t^2/2}dt =$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0}e^{-t^2/2}dt + \frac{1}{\sqrt{2\pi}}\int_{0}^{\frac{x-a}{\sigma}}e^{-t^2/2}dt.$$

Первый интеграл

$$\int_{-\infty}^{0}e^{-t^2/2}dt = \frac{1}{2}\int_{-\infty}^{+\infty}e^{-t^2/2}dt = \frac{1}{2}\sqrt{2}\int_{-\infty}^{+\infty}e^{-\left(t/\sqrt{2}\right)^2}d\left(\frac{t}{\sqrt{2}}\right) =$$

$$= \frac{\sqrt{2}}{2}\cdot\sqrt{\pi} = \sqrt{\frac{\pi}{2}}$$

(в силу четности подынтегральной функции и того, что интеграл Эйлера—Пуассона равен $\sqrt{\pi}$).

Второй интеграл с учетом **(4.29)** составляет $\dfrac{1}{2}\Phi\left(\dfrac{x-a}{\sigma}\right)$.

Итак, $F_N(x)=\frac{1}{\sqrt{2\pi}}\cdot\sqrt{\frac{\pi}{2}}+\frac{1}{2}\Phi\left(\frac{x-a}{\sigma}\right)=\frac{1}{2}+\frac{1}{2}\Phi\left(\frac{x-a}{\sigma}\right)$. ■

Геометрически функция распределения представляет собой площадь под нормальной кривой на интервале $(-\infty, x)$ (рис. 4.9). Как видим, она состоит из двух частей: первой, на интервале $(-\infty, a)$, равной 1/2, т.е. половине всей площади под нормальной кривой, и второй, на интервале (a, x), равной $\frac{1}{2}\Phi\left(\frac{x-a}{\sigma}\right)$.

Рассмотрим с в о й с т в а случайной величины, распределенной по нормальному закону.

1. *Вероятность попадания случайной величины X, распределенной по нормальному закону,* в *интервал* $[x_1, x_2]$, *равна*

$$P(x_1 \le X \le x_2)=\frac{1}{2}[\Phi(t_2)-\Phi(t_1)], \qquad (4.32)$$

где
$$t_1=\frac{x_1-a}{\sigma}, \quad t_2=\frac{x_2-a}{\sigma} \qquad (4.33)$$

□ Учитывая, что согласно **(3.19)** вероятность $P(x_1 \le X \le x_2)$ есть приращение функции распределения на отрезке $[x_1, x_2]$, и формулу **(4.30)**, получим

$$P(x_1 \le X \le x_2)=F(x_2)-F(x_1)=$$
$$=\left[\frac{1}{2}+\frac{1}{2}\Phi\left(\frac{x_2-a}{\sigma}\right)\right]-\left[\frac{1}{2}+\frac{1}{2}\Phi\left(\frac{x_1-a}{\sigma}\right)\right]=\frac{1}{2}[\Phi(t_2)-\Phi(t_1)],$$

где t_1 и t_2 определяются по формуле **(4.33)** (рис. 4.10). ■

2. *Вероятность того, что отклонение случайной величины X, распределенной по нормальному закону, от математического ожидания a не превысит величину* $\Delta > 0$ (*по абсолютной величине*), *равна*

$$P(|X-a| \le \Delta)=\Phi(t), \qquad (4.34)$$

где
$$t=\frac{\Delta}{\sigma}. \qquad (4.35)$$

□ $P(|X-a| \le \Delta)=P(a-\Delta \le X \le a+\Delta)$ Учитывая **(4.32)** и **(4.33)**, а также свойство нечетности функции Лапласа, получим

$$P(|X-a| \le \Delta)=\frac{1}{2}\left[\Phi\left(\frac{(a+\Delta)-a}{\sigma}\right)-\Phi\left(\frac{(a-\Delta)-a}{\sigma}\right)\right]=$$
$$=\frac{1}{2}\left[\Phi\left(\frac{\Delta}{\sigma}\right)-\Phi\left(-\frac{\Delta}{\sigma}\right)\right]=\frac{1}{2}\left[\Phi\left(\frac{\Delta}{\sigma}\right)+\Phi\left(\frac{\Delta}{\sigma}\right)\right]=\Phi\left(\frac{\Delta}{\sigma}\right)=\Phi(t),$$

где $t=\Delta/\sigma$ (рис. 4.11). ■

На рис. 4.10 и 4.11 приведена геометрическая интерпретация свойств нормального закона[1].

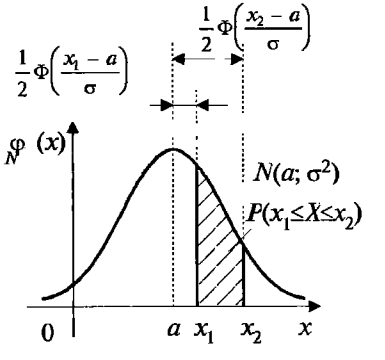


Рис. 4.10

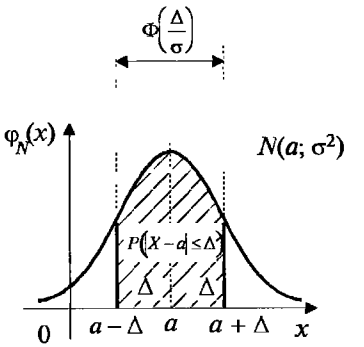


Рис. 4.11

З а м е ч а н и е. Рассмотренная в гл. 2 приближенная интегральная формула Муавра—Лапласа **(2.10)** следует из свойства **(4.32)** нормально распределенной случайной величины при $x_1=a$, $x_2=b$, $a=np$ и $\sigma_x=\sqrt{npq}$, так как биномиальный закон распределения случайной величины $X=m$ с параметрами n и p, для которого получена эта формула, при $n \to \infty$ стремится к нормальному закону (см. гл. 6).

Аналогично и следствия **(2.13)**, **(2.14)** и **(2.16)** интегральной формулы Муавра—Лапласа для числа $X=m$ появления события в n независимых испытаниях и его частости m/n вытекают из свойств **(4.32)** и **(4.34)** нормального закона.

[1] Стрелками на рис. 4.10—4.12 отмечены условно п л о щ а д и соответствующих фигур под нормальной кривой.

Вычислим по формуле **(4.34)** вероятности $P(|X-a|\le\Delta)$ при различных значениях Δ (используем табл. II приложений). Получим при

$\Delta=\sigma$ $\quad P(|X-a|\le\sigma)=\Phi(1)=0{,}6827;$

$\Delta=2\sigma$ $\quad P(|X-a|\le 2\sigma)=\Phi(2)=0{,}9545;$

$\Delta=3\sigma$ $\quad P(|X-a|\le 3\sigma)=\Phi(3)=0{,}9973$

(рис. 4.12).

Отсюда вытекает **«правило трех сигм»**:

Если случайная величина X имеет нормальный закон распределения с параметрами a и σ^2, т.е. $N(a;\ \sigma^2)$, то практически достоверно, что ее значения заключены в интервале $(a-3\sigma,\ a+3\sigma)$.

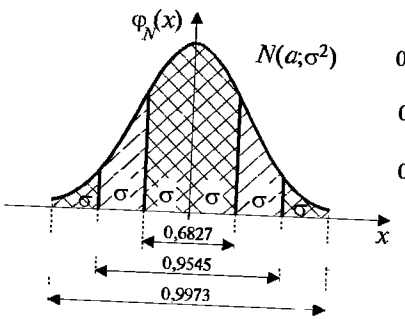


Рис. 4.12

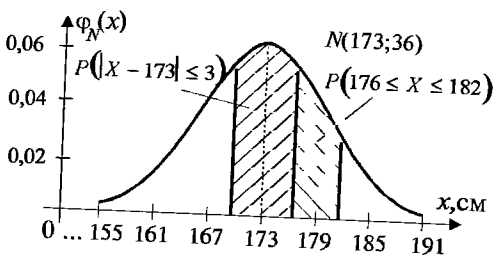


Рис. 4.13

Нарушение «правила трех сигм», т.е. отклонение нормально распределенной случайной величины X больше, чем на 3σ (по абсолютной величине), является событием практически невозможным, так как его вероятность весьма мала:

$$P(|X-a|>3\sigma)=1-P(|X-a|\le 3\sigma)=1-0{,}9973=0{,}0027.$$

Найдем **коэффициент асимметрии и эксцесс** случайной величины X, распределенной по нормальному закону.

Очевидно, в силу симметрии нормальной кривой относительно вертикальной прямой $x=a$, проходящей через центр распределения $a=M(X)$, *коэффициент асимметрии нормального распределения* $A=0$.

Эксцесс нормально распределенной случайной величины X найдем по формуле **(3.37)**, т.е.

$$E=\frac{\mu_4}{\sigma^4}-3=\frac{3\sigma^4}{\sigma^4}-3=0,$$

где учли, что центральный момент 4-го порядка, найденный по формуле **(3.30)** с учетом **(4.26)**, т.е.

$$\mu_4=\int_{-\infty}^{+\infty}(x-a)^4\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}dx=3\sigma^4$$

(вычисление интеграла опускаем).

Таким образом, *эксцесс нормального распределения равен нулю*, и крутость других распределений определяется по отношению к нормальному (об этом мы уже упоминали в § 3.7).

▷ **Пример 4.9.** Полагая, что рост мужчин определенной возрастной группы есть нормально распределенная случайная величинах X с параметрами $a=173$ и $\sigma^2=36$, найти:

1. а) выражение плотности вероятности и функции распределения случайной величины X; б) доли костюмов 4-го роста (176—182 см) и 3-го роста (170—176 см), которые нужно предусмотреть в общем объеме производства для данной возрастной группы; в) квантиль $x_{0,7}$ и 10%-ную точку случайной величины X.

2. Сформулировать «правило трех сигм» для случайной величины X.

Р е ш е н и е. 1. а) По формулам **(4.26)** и **(4.30)** запишем

$$\varphi_N(x)=\frac{1}{6\sqrt{2\pi}}e^{-\frac{(x-173)^2}{2\cdot 36}};$$

$$F_N(x)=\frac{1}{6\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(x-173)^2}{2\cdot 36}}dx=\frac{1}{2}+\frac{1}{2}\Phi\left(\frac{x-173}{6}\right).$$

б) Доля костюмов 4-го роста (176—182 см) в общем объеме производства определится по формуле **(4.32)** как вероятность[1]

$$P(176\le X\le 182)=\frac{1}{2}[\Phi(t_2)-\Phi(t_1)]=\frac{1}{2}[\Phi(1{,}50)-\Phi(0{,}50)]=$$

$$=\frac{1}{2}(0{,}8664-0{,}3829)=0{,}2418$$

(рис. 4.13), так как по **(4.33)**

$$t_1=\frac{176-173}{6}=0{,}50,\quad t_2=\frac{182-173}{6}=1{,}50.$$

[1] Значения функции Лапласа $\Phi(x)$ определяем по табл. II приложений.

Долю костюмов 3-го роста (170—176 см) можно было определить аналогично по формуле **(4.32)**, но проще это сделать по формуле **(4.34)**, если учесть, что данный интервал симметричен относительно математического ожидания $a = M(X) = 173$, т.е. неравенство $170 \le X \le 176$ равносильно неравенству $|X - 173| \le 3$:

$$P(170 \le X \le 176) = P(|X - 173| \le 3) = \Phi\left(\frac{3}{6}\right) = \Phi(0{,}50) = 0{,}3829$$

(рис. 4.13).

в) Квантиль $x_{0,7}$ (см. § 3.7) случайной величины X найдем из уравнения **(3.29)** с учетом **(4.30)**:

$$F(x_{0,7}) = \frac{1}{2} + \frac{1}{2}\Phi\left(\frac{x_{0,7} - 173}{6}\right) = 0{,}7,$$

откуда

$$\Phi\left(\frac{x_{0,7} - 173}{6}\right) = \Phi(t) = 0{,}4.$$

По табл. II приложений находим $t = 0{,}524$ и

$$x_{0,7} = 6t + 173 = 6 \cdot 0{,}524 + 173 \approx 176 \text{ (см)}.$$

Это означает, что 70% мужчин данной возрастной группы имеют рост до 176 см.

10%-ная точка — это квантиль $x_{0,9} = 181$ см (находится аналогично), т.е. 10% мужчин имеют рост не менее 181 см.

2. Практически достоверно, что рост мужчин данной возрастной группы заключен в границах от $a - 3\sigma = 173 - 3 \cdot 6 = 155$ до $a + 3\sigma = 173 + 3 \cdot 6 = 191$ (см), т.е. $155 \le X \le 191$ (см). ▶

В силу особенностей нормального закона распределения, отмеченных в начале параграфа (и в гл. 6), он занимает центральное место в теории и практике вероятностно-статистических методов. Большое теоретическое значение нормального закона состоит в том, что с его помощью получен ряд важных распределений, рассматриваемых ниже.

4.8. Логарифмически-нормальное распределение

О п р е д е л е н и е. *Непрерывная случайная величина X имеет **логарифмически-нормальное** (сокращенно **логнормальное распределение**), если ее логарифм подчинен нормальному закону.*

Так как при $x > 0$ неравенства $X < x$ и $\ln X < \ln x$ равносильны, то функция распределения логнормального распределения совпадает с функцией нормального распределения для случайной величины $\ln X$, т.е. в соответствии с **(4.31)**

$$F(x) = P(X < x) = P(\ln X < \ln x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\ln x} e^{-\frac{(t - \ln a)^2}{2\sigma^2}} dt. \qquad \textbf{(4.36)}$$

Дифференцируя **(4.36)** по x, получим выражение плотности вероятности для логнормального распределения

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}\, x} e^{-\frac{(\ln x - \ln a)^2}{2\sigma^2}} \qquad \textbf{(4.37)}$$

(рис. 4.14).

Можно доказать, что числовые характеристики случайной величины X, распределенной по логнормальному закону **(4.37)**, имеют вид: математическое ожидание $M(X) = ae^{\sigma^2/2}$, дисперсия $D(X) = a^2 e^{\sigma^2}\left(e^{\sigma^2} - 1\right)$, мода $Mo(X) = ae^{-\sigma^2}$, медиана $Me(X) = a$.

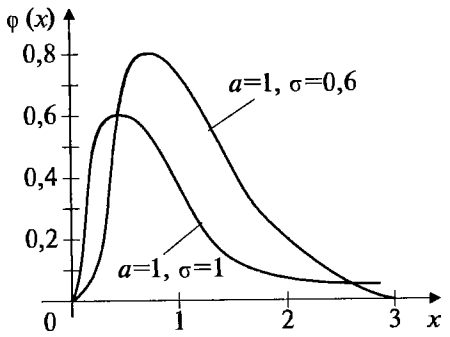


Рис. 4.14

Очевидно, чем меньше σ, тем ближе друг к другу значения моды, медианы и математического ожидания, а кривая распределения — ближе к симметрии. Если в нормальном законе параметр a выступает в качестве среднего значения случайной величины, то в логнормальном **(4.37)** — в качестве медианы.

Логнормальное распределение используется для описания распределения доходов, банковских вкладов, цен активов, месячной заработной платы, посевных площадей под разные культуры, долговечности изделий в режиме износа и старения и др.

▷ **Пример 4.10.** Проведенное исследование показало, что вклады населения в данном банке могут быть описаны случайной величиной X, распределенной по логнормальному закону **(4.37)** с параметрами $a = 530$, $\sigma^2 = 0{,}64$.

Найти: а) средний размер вклада; б) долю вкладчиков, размер вклада которых составляет не менее 1000 ден. ед.; в) моду и медиану случайной величины X и пояснить их смысл.

Р е ш е н и е. а) Найдем средний размер вклада, т.е.

$$M(X) = ae^{\sigma^2/2} = 530e^{0,64/2} = 730 \text{ (ден. ед.)}.$$

б) Доля вкладчиков, размер вклада которых составляет не менее 1000 ден. ед., есть

$$P(X \geq 1000) = 1 - P(X < 1000) = 1 - F(1000).$$

При определении $F(1000)$ воспользуемся тем, что функция логнормального распределения случайной величины X совпадает с функцией нормального распределения случайной величины $\ln X$, т.е. с учетом **(4.30)** имеем:

$$F(x) = \frac{1}{2} + \frac{1}{2}\Phi\left(\frac{\ln x - \ln a}{\sigma}\right) \text{ и}$$

$$F(1000) = \frac{1}{2} + \frac{1}{2}\Phi\left(\frac{\ln 1000 - \ln 530}{\sqrt{0,64}}\right) = \frac{1}{2} + \frac{1}{2}\Phi(0,79) =$$

$$= \frac{1}{2} + \frac{1}{2} \cdot 0,5705 = 0,785.$$

Теперь $P(X \geq 1000) = 1 - 0,785 = 0,215$ (рис. 4.15).

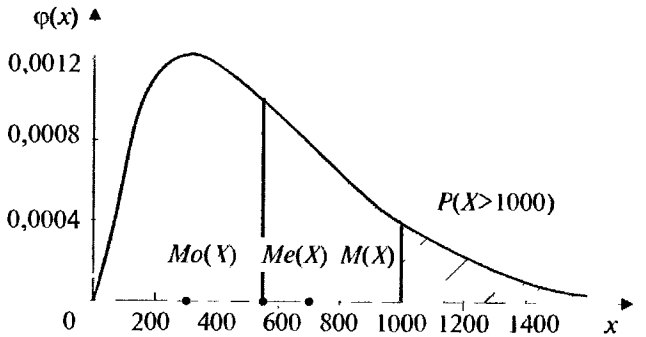


Рис. 4.15

в) Вычислим моду случайной величины X:

$Mo(X) = ae^{-\sigma^2} = 530e^{-0,64} \approx 280$, т.е. наиболее часто встречающийся банковский вклад равен 280 ден. ед. (точнее, наиболее часто встречающийся элементарный интервал с центром 280 ден. ед., т.е. интервал $(280 - \Delta, 280 + \Delta)$ ден. ед.).

Если исходить из вероятностного смысла параметра a логнормального распределения, то медиана $Me(X)=a=530$, т.е. половина вкладчиков имеют вклады до 530 ден. ед., а другая половина — сверх 530 ден. ед. ▶

4.9. Распределение некоторых случайных величин, представляющих функции нормальных величин

Ниже рассматривается несколько основных законов, составляющих необходимый аппарат для построения в дальнейшем статистических критериев и оценок, применяемых в математической статистике.

χ^2-распределение.

О п р е д е л е н и е. ***Распределением χ^2 (хи-квадрат) с k степенями свободы*** *называется распределение суммы квадратов k независимых случайных величин, распределенных по стандартному нормальному закону*, т.е.

$$\chi^2 = \sum_{i=1}^{k} Z_i^2, \qquad \textbf{(4.38)}$$

где Z_i $(i = 1, 2,...,k)$ имеет нормальное распределение $N(0;1)$.

Плотность вероятности χ^2-распределения имеет вид:

$$\varphi(x) = \begin{cases} \dfrac{1}{2^{\frac{k}{2}} \cdot \Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} & \text{при } x \geq 0, \\ 0 & \text{при } x < 0, \end{cases}$$

где $\Gamma(y) = \int_0^{+\infty} e^{-t} t^{y-1} dt$ — гамма-функция Эйлера (для целых положительных значений $\Gamma(y) = (y-1)!$).

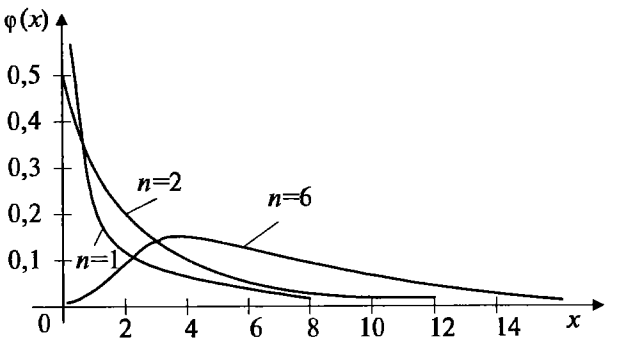


Рис. 4.16

Кривые χ^2-распределения для различных значений числа степеней свободы k приведены на рис. 4.16. Они показывают, что χ^2-распределение асимметрично, обладает положительной (правосторонней) асимметрией.

При $k>30$ распределение случайной величины $Z=\sqrt{2\chi^2}-\sqrt{2k-1}$ близко к стандартному нормальному закону, т.е. $N(0;1)$.

Распределение Стьюдента[1].

Определение. ***Распределением Стьюдента*** (или ***t*-распределением**) *называется распределение случайной величины*

$$t=\frac{Z}{\sqrt{\frac{1}{k}\chi^2}}, \qquad (4.39)$$

где Z — случайная величина, распределенная по стандартному нормальному закону, т.е. $N(0;1)$;

χ^2 — независимая от Z случайная величина, имеющая χ^2-распределение с k степенями свободы.

Плотность вероятности распределения Стьюдента имеет вид:

$$\varphi(x)=\frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{\pi k}}\left(1+\frac{x^2}{n}\right)^{-\frac{k+1}{2}},$$

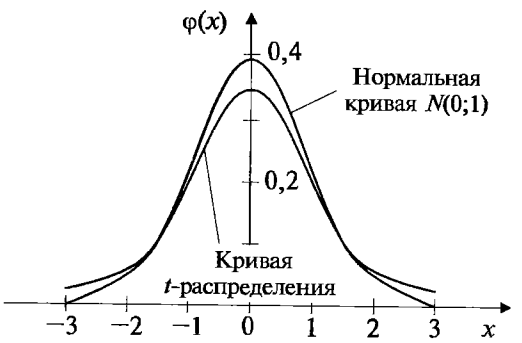


Рис. 4.17

где $\Gamma(y)$ — гамма-функция Эйлера в точке y.

На рис. 4.17 показана кривая распределения Стьюдента. Как и стандартная нормальная кривая, кривая t-распределения симметрична относительно оси ординат, но по сравнению с нормальной более пологая.

[1] Стьюдент — псевдоним английского статистика В. Госсета.

При $k\to\infty$ t-распределение приближается к нормальному. Практически уже при $k>30$ можно считать t-распределение приближенно нормальным.

Математическое ожидание случайной величины, имеющей t-распределение, в силу симметрии ее кривой распределения равно нулю, а ее дисперсия (как можно доказать) равна $k/(k-2)$, т.е. $M(t)=0$, $D(t)=\frac{k}{k-2}$.

Распределение Фишера—Снедекора.

Определение. ***Распределением Фишера—Снедекора*** (*или* ***F*-распределением**) *называется распределение случайной величины*

$$F=\frac{\frac{1}{k_1}\chi^2(k_1)}{\frac{1}{k_2}\chi^2(k_2)}, \qquad (4.40)$$

где $\chi^2(k_1)$ и $\chi^2(k_2)$ — случайные величины, имеющие χ^2-распределение соответственно с k_1 и k_2 степенями свободы.

Плотность вероятности F-распределения имеет вид:

$$\varphi(x)=\frac{\Gamma\left(\frac{k_1+k_2}{2}\right)k_1^{\frac{k_1}{2}}k_2^{\frac{k_2}{2}}}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)}x^{\frac{k_1}{2}-1}(k_1x+k_2)^{-\frac{k_1+k_2}{2}},$$

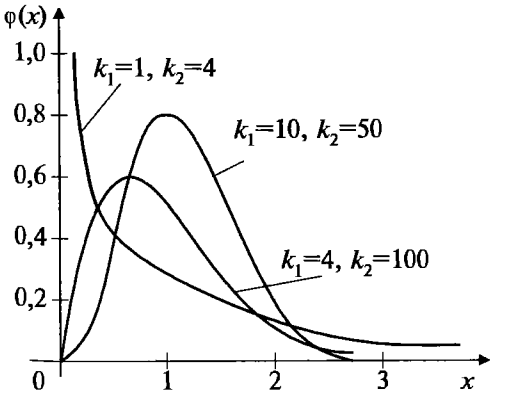


Рис. 4.18

где $\Gamma(y)$ — гамма-функция Эйлера в точке y.

На рис. 4.18 показаны кривые F-распределения при некоторых значениях числа степеней свободы k_1 и k_2. При $n\to\infty$ F-распределение приближается к нормальному закону.

Упражнения

4.11. Вероятность выигрыша по облигации займа за все время его действия равна 0,1. Составить закон распределения числа выигравших облигаций среди приобретенных 19. Найти математическое ожидание, дисперсию, среднее квадратическое отклонение и моду этой случайной величины.

4.12. По данным примера **4.11** найти математическое ожидание, дисперсию и среднее квадратическое отклонение доли (частости) выигравших облигаций среди приобретенных.

4.13. Составить функцию распределения случайной величины, имеющей биномиальный закон распределения с параметрами n и p.

4.14. Устройство состоит из 1000 элементов, работающих независимо один от другого. Вероятность отказа любого элемента в течение времени t равна 0,002. Необходимо: а) составить закон распределения отказавших за время t элементов; б) найти математическое ожидание и дисперсию этой случайной величины; в) определить вероятность того, что за время t откажет хотя бы один элемент.

4.15. Вероятность поражения цели равна 0,05. Производится стрельба по цели до первого попадания. Необходимо: а) составить закон распределения числа сделанных выстрелов; б) найти математическое ожидание и дисперсию этой случайной величины; в) определить вероятность того, что для поражения цели потребуется не менее 5 выстрелов.

4.16. В магазине имеются 20 телевизоров, из них 7 имеют дефекты. Необходимо: а) составить закон распределения числа телевизоров с дефектами среди выбранных наудачу пяти; б) найти математическое ожидание и дисперсию этой случайной величины; в) определить вероятность того, что среди выбранных нет телевизоров с дефектами.

4.17. Цена деления шкалы измерительного прибора равна 0,2. Показания прибора округляют до ближайшего целого числа. Полагая, что при отсчете ошибка округления распределена по равномерному закону, найти: 1) математическое ожидание, дисперсию и среднее квадратическое отклонение этой случайной величины; 2) вероятность того, что ошибка округления: а) меньше 0,04; б) больше 0,05.

4.18. Среднее время безотказной работы прибора равно 80 ч. Полагая, что время безотказной работы прибора имеет показательный закон распределения, найти: а) выражение его плотности вероятности и функции распределения; б) вероятность того, что в течение 100 ч прибор не выйдет из строя.

4.19. Текущая цена акции может быть смоделирована с помощью нормального закона распределения с математическим ожиданием 15 ден. ед. и средним квадратическим отклонением 0,2 ден. ед. 1. Найти вероятность того, что цена акции: а) не выше 15,3 ден. ед.; б) не ниже 15,4 ден. ед.; в) от 14,9 до 15,3 ден. ед. 2. С помощью правила трех сигм найти границы, в которых будет находиться текущая цена акции.

4.20. Цена некой ценной бумаги нормально распределена. В течение последнего года 20% рабочих дней она была ниже 88 ден. ед., а 75% — выше 90 ден. ед. Найти: а) математическое ожидание и среднее квадратическое отклонение цены ценной бумаги; б) вероятность того, что в день покупки цена будет заключена в пределах от 83 до 96 ден. ед.; в) с надежностью 0,95 определить максимальное отклонение цены ценной бумаги от среднего (прогнозного) значения (по абсолютной величине).

4.21. Коробки с конфетами упаковываются автоматически. Их средняя масса равна 540 г. Известно, что масса коробок с конфетами имеет нормальное распределение, а 5% коробок имеют массу, меньшую 500 г. Каков процент коробок, масса которых: а) менее 470 г; б) от 500 до 550 г; в) более 550 г; г) отличается от средней не более, чем на 30 г (по абсолютной величине)?

4.22. Случайная величина X имеет нормальное распределение с математическим ожиданием $a = 25$. Вероятность попадания X в интервал (10; 15) равна 0,09. Чему равна вероятность попадания X в интервал: а) (35;40); б) (30;35)?

4.23. Нормально распределенная случайная величина имеет следующую функцию распределения: $F(x)=0{,}5+ +0{,}5\,\Phi(x-1)$. Из какого интервала (1;2) или (2;6) она примет значение с большей вероятностью?

4.24. Квантиль уровня 0,15 нормально распределенной случайной величины X равен 12, а квантиль уровня 0,6

равен 16. Найти математическое ожидание и среднее квадратическое отклонение случайной величины.

4.25. 20%-ная точка нормально распределенной случайной величины равна 50, а 40%-ная точка равна 35. Найти вероятность того, что случайная величина примет значение в интервале (25;45).

4.26. Месячный доход семей можно рассматривать как случайную величину, распределенную по логнормальному закону. Полагая, что математическое ожидание этой случайной величины равно 1000 ден. ед., а среднее квадратическое отклонение 800 ден. ед., найти долю семей, имеющих доход: а) не менее 1000 ден. ед.; б) менее 500 ден. ед.

4.27. Известно, что нормально распределенная случайная величина принимает значение: а) меньшее 248 с вероятностью 0,975; б) большее 279 с вероятностью 0,005. Найти функцию распределения случайной величины X.

4.28. Случайная величина X распределена по нормальному закону с нулевым математическим ожиданием. Вероятность попадания этой случайной величины на отрезок от -1 до $+1$ равна 0,5. Найти выражения плотности вероятности и функции распределения случайной величины X.

4.29. Имеется случайная величина X, распределенная по нормальному закону с математическим ожиданием a и дисперсией σ^2. Требуется приближенно заменить нормальный закон распределения равномерным законом в интервале $(\alpha; \beta)$; границы α, β подобрать так, чтобы сохранить неизменными математическое ожидание и дисперсию случайной величины X.

4.30. Случайная величина X распределена по нормальному закону с математическим ожиданием $a = 0$. При каком значении среднего квадратического отклонения σ вероятность попадания случайной величины X в интервал (1;2) достигает максимума?

4.31. Время ремонта телевизора распределено по показательному закону с математическим ожиданием, равным 0,5 ч. Некто сдает в ремонт два телевизора, которые одновременно начинают ремонтироваться, и ждет, когда будет отремонтирован один из них. После этого с готовым телевизором он уходит. Найти закон распределения времени: а) потраченного клиентом; б) которое должен потратить клиент, если он хочет забрать сразу два телевизора.

Глава 5 Многомерные случайные величины

5.1. Понятие многомерной случайной величины и закон ее распределения

Очень часто результат испытания характеризуется не одной случайной величиной, а некоторой *системой случайных величин* $X_1, X_2,...,X_n$, которую называют также *многомерной* (*n-мерной*) *случайной величиной* или *случайным вектором* $X = (X_1, X_2,...,X_n)$.

Приведем примеры многомерных случайных величин.

1. Успеваемость выпускника вуза характеризуется системой n случайных величин $X_1, X_2,...,X_n$ — оценками по различным дисциплинам, проставленными в приложении к диплому.

2. Погода в данном месте в определенное время суток может быть охарактеризована системой случайных величин: X_1 — температура; X_2 — влажность; X_3 — давление; X_4 — скорость ветра и т.п.

В теоретико-множественной трактовке любая случайная величина X_i $(i = 1,2,...,n)$ есть функция элементарных событий ω, входящих в пространство элементарных событий $\Omega(\omega \in \Omega)$. Поэтому и *многомерная случайная величина есть функция элементарных событий* ω:

$$(X_1, X_2,..., X_n) = f(\omega),$$

т.е. каждому элементарному событию ω ставится в соответствие несколько действительных чисел $x_1, x_2,...,x_n$, которые приняли случайные величины $X_1, X_2,...,X_n$ в результате испытания. В этом случае вектор $x=(x_1, x_2,...,x_n)$ называется *реализацией* случайного вектора $X=(X_1, X_2,...,X_n)$.

Случайные величины $X_1, X_2,...,X_n$, входящие в систему, могут быть как *дискретными* (см. выше пример 1), так и *непрерывными* (пример 2).

▷ **Пример 5.1.** Подбрасывают одновременно две игральные кости; случайная величина X — сумма очков, полученных в ре-

зультате испытания; случайная величина Y — их произведение. Показать, что двумерная случайная величина (X, Y) есть функция элементарных исходов (событий) ω.

Р е ш е н и е. Множество элементарных исходов (пространство элементарных событий) состоит из 36 элементарных исходов, т.е.

$$\Omega = \{\omega_1, \omega_2, ..., \omega_{36}\} =$$
$$\{1/1,\ 1/2,...,1/6,\ 2/1,\ 2/2,...,2/6,...,6/1,\ 6/2,...,6/6\},$$

где элементарный исход, например ω_9=2/3, означает выпадение при подбрасывании первой игральной кости 2 очков и второй кости — 3 очков. Если результатом испытания является какой-нибудь элементарный исход (событие) ω_i, то случайные величины X и Y получат определенные значения; например, при ω_9=2/3 X=5, Y=6. Совокупность этих значений (X,Y) представляет, таким образом, функцию элементарных исходов (событий) ω. ▶

Геометрически двумерную (X, Y) и трехмерную (X, Y, Z) случайные величины можно изобразить случайной точкой или случайным вектором плоскости Oxy или трехмерного пространства $Oxyz$; при этом случайные величины X, Y или X, Y, Z являются *составляющими* этих векторов. В случае n-мерного пространства ($n > 3$) также говорят о случайной точке или случайном векторе этого пространства, хотя геометрическая интерпретация в этом случае теряет свою наглядность.

Наиболее полным, исчерпывающим описанием многомерной случайной величины является *закон ее распределения*. При конечном множестве возможных значений многомерной случайной величины такой закон может быть задан в форме таблицы (матрицы), содержащей всевозможные сочетания значений каждой из одномерных случайных величин, входящих в систему, и соответствующие им вероятности. Так, если рассматривается *двумерная дискретная случайная величина* (X,Y), то ее двумерное распределение можно представить в виде *таблицы* (*матрицы*) *распределения* (табл. 5.1), в каждой клетке (i,j) которой располагаются вероятности произведения событий $p_{ij} = P[(X = x_i)(Y = y_j)]$.

Таблица 5.1

x_i \ y_j	y_1	...	y_j	...	y_m	$\sum_{j=1}^{m}$
x_1	p_{11}	...	p_{1j}	...	p_{1m}	p_1
...	...	...	...	...	...	...
x_i	p_{i1}	...	p_{ij}	...	p_{im}	p_i
...	...	...	...	...	...	...
x_n	p_{n1}	...	p_{nj}	...	p_{nm}	p_n
$\sum_{i=1}^{m}$	p_1	...	p_j	...	p_m	1

Так как события $[(X = x_i)(Y = y_j)]$ (i = 1,2,...,n; j = 1,2,...,m), состоящие в том, что случайная величина X примет значение x_i, а случайная величина Y — значение y_j, несовместны и единственно возможны, т.е. образуют полную группу, то сумма их вероятностей равна единице, т.е.

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} = 1.$$

Итоговые столбец или строка таблицы распределения (X,Y) представляют соответственно распределения одномерных составляющих (x_i, p_i) или (y_j, p_j).

Действительно, распределение одномерной случайной величины X можно получить, вычислив вероятность события $X = x_i$ (i=1,2,...,n) как сумму вероятностей несовместных событий:

$$p_i = P(X = x_i) =$$
$$= P[(X = x_i)(Y = y_1) + ... + (X = x_i)(Y = y_j) + ... + (X = x_i)(Y = y_m)] =$$
$$= p_{i1} + ... + p_{ij} + ... + p_{im} = \sum_{j=1}^{m} p_{ij}.$$

Аналогично $p_j = \sum_{i=1}^{n} p_{ij}$

Таким образом, *чтобы по таблице распределения (табл. 5.1) найти вероятность того, что одномерная случайная величина примет определенное значение, надо просуммировать вероятности p_{ij} из соответствующей этому значению строки (столбца) данной таблицы.*

Если зафиксировать значение одного из аргументов, например, положить $Y = y_j$, то полученное распределение случайной

величины X называется *условным распределением* X при условии $Y = y_j$. Вероятности $p_j(x_i)$ этого распределения будут *условными вероятностями* события $X = x_i$, найденными в предположении, что событие $Y = y_j$ произошло. Из определения условной вероятности[1] **(1.34)**

$$p_j(x_i) = \frac{P\left[(X = x_i)(Y = y_j)\right]}{P(Y = y_j)} = \frac{p_{ij}}{p_j}. \qquad \textbf{(5.1)}$$

Аналогично *условное распределение случайной величины* Y при условии $X = x_i$ задается с помощью условных вероятностей

$$p_i(y_j) = \frac{P\left[(X = x_i)(Y = y_j)\right]}{P(X = x_i)} = \frac{p_{ij}}{p_i}. \qquad \textbf{(5.2)}$$

▷ **Пример 5.2.** Закон распределения дискретной двумерной случайной величины (X, Y) задан в табл. 5.2.

Таблица 5.2

x_i \ y_j	−1	0	1	2
1	0,10	0,25	0,30	0,15
2	0,10	0,05	0,00	0,05

Найти: а) законы распределения одномерных случайных величин X и Y; б) условные законы распределения случайной величины X при условии $Y = 2$ и случайной величины Y при условии $X = 1$; в) вычислить $P(Y < X)$.

Р е ш е н и е а) Случайная величина X может принимать значения:

$X = 1$ с вероятностью p_1=0,10+0,25+0,30+0,15=0,8;

$X = 2$ с вероятностью p_2=0,10+0,05+0,00+0,05=0,2,

[1] Для условных вероятностей используются также обозначения $P(x_i | y_j)$, $P(y_j | x_i)$.

т.е. ее закон распределения

X:	x_i	1	2
	p_i	0,8	0,2

Аналогично закон распределения

Y:	y_j	−1	0	1	2
	p_j	0,2	0,3	0,3	0,2

б) Условный закон распределения X при условии, что $Y = 2$, получим, если вероятности p_{ij}, стоящие в последнем столбце табл. 5.2, разделим на их сумму, т.е. $p(Y = 2) = 0,2$. Получим

$X_{Y=2}$:	x_i	1	2
	$p_j(x_i)$	0,75	0,25

Аналогично для получения условного закона распределения Y при условии $X = 1$ вероятности p_{ij}, стоящие в первой строке табл. 5.2, делим на их сумму, т.е. на $p(X = 1) = 0,8$. Получим

$Y_{X=1}$:	y_j	−1	0	1	2
	$p_i(y_j)$	0,125	0,3125	0,375	0,1875

в) Для нахождения вероятностей $P(Y < X)$ складываем вероятности событий p_{ij} из табл. 5.2, для которых $y_j < x_i$.

Получим

$$P(Y < X)=0,10+0,25+0,10+0,05+0,00=0,5. \blacktriangleright$$

5.2. Функция распределения многомерной случайной величины

О п р е д е л е н и е. ***Функцией распределения n-мерной случайной величины*** $(X_1, X_2, \ldots, X_n)$ *называется функция* $F(x_1, x_2, \ldots, x_n)$, *выражающая вероятность совместного выполнения n неравенств*[1] $X_1 < x_1$, $X_2 < x_2, \ldots, X_n < x_n$, т.е.

[1] Функцию $F(x_1, x_2, \ldots, x_n)$ называют также *совместной функцией распределения* случайных величин $X_1, X_2, \ldots, X_n$.

$$F(x_1, x_2, \ldots, x_n) = P(X_1 < x_1,\ X_2 < x_2, \ldots, X_n < x_n). \quad (5.3)$$

В двумерном случае[1] для случайной величины (X, Y) функция распределения $F(x,y)$ определится равенством:

$$F(x, y) = P(X < x,\ Y < y). \quad (5.4)$$

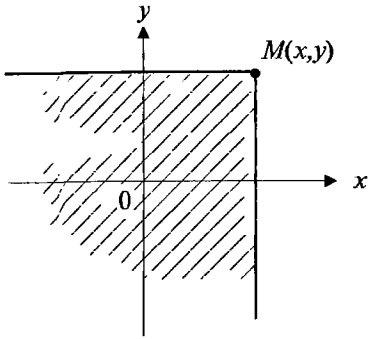


Рис. 5.1

Геометрически функция распределения $F(x,y)$ означает вероятность попадания случайной точки (X, Y) в заштрихованную область — бесконечный квадрант, лежащий левее и ниже точки $M(x,y)$ (рис. 5.1). Правая и верхняя границы области в квадрант не включаются — это означает, что функция распределения непрерывна слева по каждому из аргументов.

В случае дискретной двумерной случайной величины ее функция распределения определяется по формуле:

$$F(x, y) = \sum_i \sum_j p_{ij}, \quad (5.5)$$

где суммирование вероятностей распространяется на все i, для которых $x_i < x$, и все j, для которых $y_j < y$.

Отметим свойства функции распределения двумерной случайной величины, аналогичные свойствам функции распределения одномерной случайной величины.

1. *Функция распределения $F(x, y)$ есть неотрицательная функция, заключенная между нулем и единицей*, т.е.

$$0 \le F(x, y) \le 1. \quad (5.6)$$

□ Утверждение следует из того, что $F(x, y)$ есть вероятность. ■

[1] В основном будем вести изложение для двумерной (n=2) случайной величины; при этом практически все понятия и утверждения, сформулированные для n=2, могут быть перенесены и на случай n>2. Однако рассмотрение двумерной случайной величины позволяет сделать изложение наглядным и менее громоздким.

2. *Функция распределения $F(x, y)$ есть неубывающая функция по каждому из аргументов*, т.е.

$$\text{при } x_2 > x_1 \quad F(x_2, y) \ge F(x_1, y),$$

$$\text{при } y_2 > y_1 \quad F(x, y_2) \ge F(x, y_1). \quad (5.7)$$

□ Так как при увеличении какого-либо аргумента заштрихованная область на рис. 5.1 увеличивается, то вероятность попадания в него случайной точки (X, Y), т.е. функция распределения $F(x, y)$, уменьшиться не может. ■

3. *Если хотя бы один из аргументов обращается в $-\infty$, функция распределения $F(x, y)$ равна нулю*, т.е.

$$F(x, -\infty) = F(-\infty, y) = F(-\infty, -\infty) = 0. \quad (5.8)$$

□ Функция распределения $F(x, y)$ в отмеченных случаях равна нулю, так как события $X < -\infty$, $Y < -\infty$ и их произведение представляют невозможные события. ■

4. *Если один из аргументов обращается в $+\infty$, функция распределения $F(x, y)$ становится равной функции распределения случайной величины, соответствующей другому аргументу*:

$$F(x, +\infty) = F_1(x),$$

$$F(+\infty, y) = F_2(y), \quad (5.9)$$

где $F_1(x)$ и $F_2(y)$ — *функции распределения случайных величин* X и Y, т.е.

$$F_1(x) = P(X < x), \quad F_2(y) = P(Y < y).$$

□ Произведение события $(X < x)$ и достоверного события $(Y < +\infty)$ есть само событие $(X < x)$, следовательно, $F(x, +\infty) = P(X < x) = F_1(x)$. Аналогично можно показать, что $F(+\infty, y) = F_2(y)$. ■

5. *Если оба аргумента равны* $+\infty$, *то функция распределения равна единице*:

$$F(+\infty,+\infty)=1.$$

□ $F(+\infty,+\infty)=1$ следует из того, что совместное осуществление достоверных событий $(X<+\infty)$, $(Y<+\infty)$ есть событие достоверное. ■

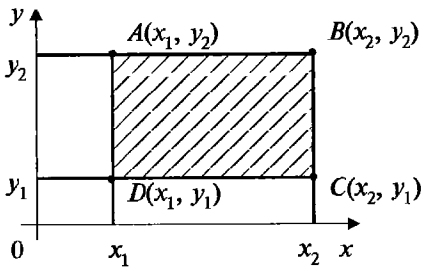


Рис. 5.2

Геометрически функция распределения есть некоторая поверхность, обладающая указанными свойствами. Для дискретной случайной двумерной величины (X,Y) ее функция распределения представляет собой некоторую ступенчатую поверхность, ступени которой соответствуют скачкам функции $F(x, y)$.

Зная функцию распределения $F(x, y)$, можно найти вероятность попадания случайной точки (X,Y) в пределы прямоугольника $ABCD$ (рис. 5.2), т.е. $P\left[(x_1 \le X < x_2)(y_1 \le Y < y_2)\right]$. Так как эта вероятность равна вероятности попадания в бесконечный квадрант с вершиной $B(x_2, y_2)$ минус вероятности попадания в квадранты с вершинами соответственно в точках $A(x_1, y_2)$ и $C(x_2, y_1)$ плюс вероятность попадания в квадрант с вершиной в точке $D(x_1, y_1)$ (ибо эта вероятность вычиталась дважды), то

$$P\left[(x_1 \le X < x_2)(y_1 \le Y < y_2)\right] = F(x_2,y_2)-F(x_1,y_2)-F(x_2,y_1)+F(x_1,y_1). \qquad (5.10)$$

5.3. Плотность вероятности двумерной случайной величины

Определение. *Двумерная случайная величина* (X,Y) *называется* ***непрерывной****, если ее функция распределения* $F(x,y)$ *— непрерывная функция, дифференцируемая по каждому из аргументов, и существует вторая смешанная производная* $F''_{xy}(x,y)$.

Аналогично одномерному случаю вероятность пары отдельно взятых значений двумерной непрерывной случайной величины равна нулю, т.е. $P(X = x_1,\ Y = y_1)=0$. Это вытекает непосредственно из формулы **(5.10)** при $x_2 \to x_1$, $y_2 \to y_1$ с учетом непрерывности функции распределения $F(x,y)$.

Для двумерной случайной величины, так же как и для одномерной, вводится понятие плотности вероятности.

Найдем вероятность попадания случайной точки (X,Y) в прямоугольник со сторонами Δx и Δy, т.е.

$$\mathcal{P} = P\left[(x \le X < x+\Delta x)(y \le Y < y+\Delta y)\right].$$

Полагая в формуле **(5.10)** $x_1=x$, $x_2=x+\Delta x, y_1 = y$, $y_2 = y+\Delta y$, получим, что эта вероятность

$$\mathcal{P} = \left[F(x+\Delta x, y+\Delta y)-F(x, y+\Delta y)\right]- \left[F(x+\Delta x, y)-F(x,y)\right] \qquad (5.11)$$

Средняя плотность вероятности в данном прямоугольнике равна отношению вероятности $\mathcal{P}$ к площади прямоугольника $\Delta x\Delta y$, т.е.

$$\varphi_{ср} = \frac{\mathcal{P}\left[(x \le X < x+\Delta x)(y \le Y < y+\Delta y)\right]}{\Delta x \cdot \Delta y}. \qquad (5.12)$$

Будем неограниченно уменьшать стороны прямоугольника, т.е. $\Delta x \to 0, \Delta y \to 0$. Тогда, учитывая **(5.11)**, найдем

$$\lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}} \varphi_{-p} = \lim_{\Delta y\to 0}\frac{1}{\Delta y}\left[\lim_{\Delta x\to 0}\frac{F(x+\Delta x, y+\Delta y)-F(x, y+\Delta y)}{\Delta x} - \lim_{\Delta x\to 0}\frac{F(x+\Delta x)-F(x,y)}{\Delta x}\right]. \qquad (5.13)$$

Так как функция $F(x,y)$ непрерывна и дифференцируема по каждому аргументу, то выражение **(5.13)** примет вид:

$$\lim_{\Delta x\to 0}\varphi_{-p} = \lim_{\Delta y\to 0}\left[\frac{F'_x(x,y+\Delta y)-F'_x(x,y)}{\Delta y}\right] = \left[F'_x(x,y)\right]'_y = F''_{xy}(x,y). \qquad (5.14)$$

Определение. ***Плотностью вероятности*** **(*плотностью распределения* или *совместной плотностью*)** *непрерывной двумер-*

ной случайной величины (X,Y) называется вторая смешанная частная производная ее функции распределения, т.е.

$$\varphi(x,y)=\frac{\partial^2 F(x,y)}{\partial x\partial y}=F''_{xy}(x,y). \qquad (5.15)$$

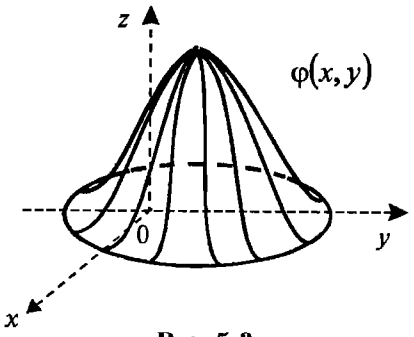


Рис. 5.3

Геометрически плотность вероятности двумерной случайной величины (X,Y) представляет собой *поверхность распределения* в пространстве $Oxyz$ (рис. 5.3).

Плотность вероятности $\varphi(x,y)$ обладает с в о й с т в а м и, аналогичными свойствам плотности вероятности одномерной случайной величины:

1. *Плотность вероятности двумерной случайной величины есть неотрицательная функция*, т.е.

$$\varphi(x,y)\geq 0.$$

□ Свойство вытекает из определения плотности вероятности как предела отношения **(5.13)** двух неотрицательных величин, ибо функция распределения $F(x,y)$ — неубывающая функция по каждому аргументу. ■

2. *Вероятность попадания непрерывной двумерной величины (X,Y) в область D равна*

$$P[(X,Y)\in D]=\iint_D \varphi(x,y)\,dx\,dy. \qquad (5.16)$$

□ Поясним геометрически формулу **(5.16)**.

Подобно тому, как в гл. 4 для одномерной случайной величины X введено понятие «элемент вероятности», равный $\varphi(x)dx$, для двумерной случайной величины (X,Y) вводится также понятие «элемент вероятности», равный $\varphi(x,y)\,dx\,dy$. Он представляет (с точностью до бесконечно малых более высоких порядков) вероятность попадания случайной точки (X,Y) в элементарный прямоугольник со сторонами dx и dy (рис. 5.4).

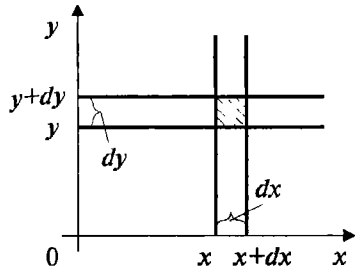


Рис. 5.4

Эта вероятность приближенно равна объему элементарного параллелепипеда с высотой $\varphi(x,y)$, опирающегося на элементарный прямоугольник со сторонами dx и dy.

Если вероятность попадания одномерной случайной величины на отрезок $[a,b]$ геометрически выражалась площадью фигуры, ограниченной сверху кривой распределения $\varphi(x)$ и опирающейся на отрезок $[a,b]$, и аналитически выражалась интегралом $\int_a^b \varphi(x)\,dx$, то *вероятность попадания двумерной случайной величины в область D на плоскости Oxy геометрически изображается объемом цилиндрического тела, ограниченного сверху поверхностью распределения $\varphi(x,y)$ и опирающегося на область D*, а аналитически — двойным интегралом **(5.16)**. ■

3. *Функция распределения непрерывной двумерной случайной величины может быть выражена через ее плотность вероятности $\varphi(x,y)$ по формуле*:

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}\varphi(x,y)\,dx\,dy. \qquad (5.17)$$

□ Функция распределения $F(x,y)$ есть вероятность попадания в бесконечный квадрант D, который можно рассматривать как прямоугольник, ограниченный абсциссами $-\infty$ и x и ординатами $-\infty$ и y. Поэтому в соответствии с **(5.16)**

$$F(x,y)=\iint_D \varphi(x,y)\,dx\,dy=\int_{-\infty}^{x}\int_{-\infty}^{y}\varphi(x,y)\,dx\,dy. \ ■$$

4. *Двойной несобственный интеграл в бесконечных пределах от плотности вероятности двумерной случайной величины равен единице*:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\varphi(x,y)\,dx\,dy=1. \qquad (5.18)$$

□ Несобственный интеграл **(5.18)** есть вероятность попадания во всю плоскость Oxy, т.е. вероятность достоверного события, равная 1. Это означает, что *полный объем тела, ограниченного поверхностью распределения и плоскостью Oxy, равен* 1. ■

Зная плотность вероятности двумерной случайной величины (X, Y), можно найти функции распределения и плотности вероятностей ее одномерных составляющих X и Y.

Так как в соответствии с **(5.9)** $F(x,+\infty)=F_1(x)$ и $F(+\infty, y)=F_2(y)$, то взяв в формулах **(5.17)** соответственно $y=+\infty$ и $x=+\infty$, получим функции распределения одномерных случайных величин X и Y:

$$F_1(x)=\int_{-\infty}^{x}\int_{-\infty}^{+\infty}\varphi(x,y)\,dx\,dy,\qquad F_2(y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{y}\varphi(x,\,y)\,dx\,dy. \tag{5.19}$$

Дифференцируя функции распределения $F_1(x)$ и $F_2(y)$ соответственно по аргументам x и y, получим плотности вероятности одномерных случайных величин X и Y:

$$\varphi_1(x)=\int_{-\infty}^{+\infty}\varphi(x,y)\,dy,\qquad \varphi_2(y)=\int_{-\infty}^{+\infty}\varphi(x,y)\,dx, \tag{5.20}$$

т.е. *несобственный интеграл в бесконечных пределах от совместной плотности $\varphi(x,y)$ двумерной случайной величины по аргументу x дает плотность вероятности $\varphi_2(y)$, а по аргументу y — плотность вероятности $\varphi_1(x)$.*

Замечание. Если имеется кривая распределения $\varphi(x)$ одномерной случайной величины X, то конкретное значение ее плотности вероятности в данной точке x определяется геометрически ординатой кривой $\varphi(x)$. Если имеется поверхность распределения $\varphi(x,y)$ двумерной случайной величины (X,Y), то конкретное значение ее совместной плотности в данной точке (x,y) определяется геометрически аппликатой поверхности $\varphi(x,y)$. В этом случае конкретное значение плотности вероятности $\varphi_1(x)$ одномерной составляющей X в данной точке x, в соответствии с **(5.20)**, определится геометрически площадью сечения поверхности $\varphi(x,y)$ плоскостью $X=x$, параллельной координатной плоскости Oyz и отсекающей

на оси Ox отрезок x. Аналогично конкретное значение плотности $\varphi_2(y)$ одномерной составляющей Y в данной точке y есть площадь сечения поверхности $\varphi(x,y)$ плоскостью $Y=y$, параллельной координатной плоскости Oxz и отсекающей на оси Oy отрезок y (см. рис. 5.9, на котором значение $\varphi_2(y)$ при данном y представляет площадь сечения, равную s).

▷ **Пример 5.3.** Двумерная случайная величина распределена равномерно в круге радиуса $R=1$ (рис. 5.5). Определить: а) выражение совместной плотности и функции распределения двумерной случайной величины (X,Y); б) плотности вероятности и функции распределения одномерных составляющих X и Y; в) вероятность того, что расстояние от точки (X,Y) до начала координат будет меньше 1/3.

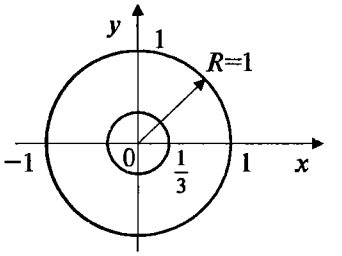


Рис. 5.5

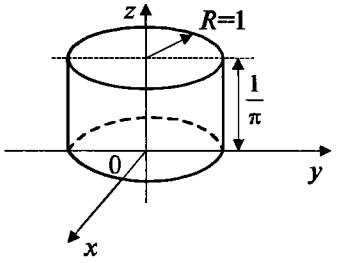


Рис. 5.6

Р е ш е н и е. а) По условию $\varphi(x,y)=\begin{cases}C \text{ при } x^2+y^2\le 1,\\ 0 \text{ при } x^2+y^2>1.\end{cases}$

Постоянную C можно найти из соотношения **(5.18)**:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\varphi(x,y)\,dx\,dy=\int_{-1}^{1}\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}C\,dx\,dy=1.$$

Проще это сделать, исходя из геометрического смысла соотношения **(5.18)**, означающего, что объем тела, ограниченного поверхностью распределения $\varphi(x,y)$ и плоскостью Oxy, равен 1. В данном случае, это объем цилиндра с площадью основания $\pi R^2=\pi\cdot 1^2=\pi$ и высотой C (рис. 5.6), равный $\pi\cdot C=1$, откуда $C=1/\pi$. Следовательно,

$$\varphi(x,y)=\begin{cases}1/\pi & \text{при } x^2+y^2\le 1,\\ 0 & \text{при } x^2+y^2>1.\end{cases}$$

Найдем функцию распределения $F(x,y)$ по формуле **(5.17)**:

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}\varphi(x,y)\,dx\,dy=\frac{1}{\pi}\int_{-\infty}^{x}dx\int_{-\infty}^{y}dy. \qquad (5.21)$$

Очевидно, что этот интеграл с точностью до множителя $1/\pi$ совпадает с площадью области D — области пересечения круга $x^2+y^2\le 1$ с бесконечным квадрантом левее и ниже точки $M(x,y)$ (рис. 5.7).

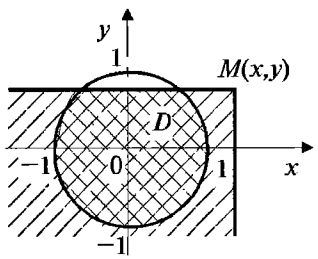


Рис. 5.7

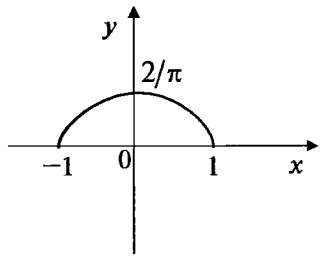


Рис. 5.8

Опустим расчеты интеграла **(5.21)** для различных x и y, предоставив их читателю, но отметим очевидное, что при $x\le -1$, $-\infty<y<+\infty$ или при $-\infty<x<+\infty$, $y\le -1$ $F(x,y)=0$, так как в этом случае область D — пустая, а при $x>1$, $y>1$ $F(x,y)=1$, так как при этом область D полностью совпадает с кругом $x^2+y^2\le 1$, на котором совместная плотность $\varphi(x,y)$ отлична от нуля.

б) Найдем функции распределения одномерных составляющих X и Y. По формуле **(5.19)** при $-1<x\le 1$

$$F_1(x)=\int_{-\infty}^{x}\int_{-\infty}^{+\infty}\varphi(x,y)\,dx\,dy=\int_{-1}^{x}\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}\frac{1}{\pi}\,dx\,dy=$$
$$=\frac{1}{\pi}\int_{-1}^{x}\left(y\Big|_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}\right)dx=\frac{1}{\pi}\int_{-1}^{x}2\sqrt{1-x^2}\,dx=$$
$$=\frac{1}{\pi}\left[\left(x\sqrt{1-x^2}\right)+\arcsin x\Big|_{-1}^{x}\right]=\frac{1}{2}+\frac{1}{\pi}\left(x\sqrt{1-x^2}+\arcsin x\right)^{1}.$$

[1] Можно показать, что для функции $2\sqrt{1-x^2}$ первообразная есть $x\sqrt{1-x^2}+\arcsin x$.

Итак,

$$F_1(x)=\begin{cases}0 & \text{при} \quad x\le -1,\\ \frac{1}{2}+\frac{1}{\pi}(x\sqrt{1-x^2}+\arcsin x) & \text{при} \quad -1<x\le 1,\\ 1 & \text{при} \quad x>1.\end{cases}$$

Аналогично

$$F_2(y)=\begin{cases}0 & \text{при} \quad y\le -1,\\ \frac{1}{2}+\frac{1}{\pi}(y\sqrt{1-y^2}+\arcsin y) & \text{при} \quad -1<y\le 1,\\ 1 & \text{при} \quad y>1.\end{cases}$$

Найдем плотности вероятности одномерных составляющих X и Y. По формуле **(5.20)**

$$\varphi_1(x)=\int_{-\infty}^{+\infty}\varphi(x,y)dy=\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}\frac{1}{\pi}dy=\frac{2}{\pi}\sqrt{1-x^2}\,(-1\le x\le 1).$$

График плотности $\varphi_1(x)$ показан на рис. 5.8.

Аналогично

$$\varphi_2(y)=\int_{-\infty}^{+\infty}\varphi(x,y)dx=\frac{2}{\pi}\sqrt{1-y^2}\,(-1\le y\le 1).$$

в) Искомую вероятность $P\left(\sqrt{X^2+Y^2}<\frac{1}{3}\right)=P\left(X^2+Y^2<\frac{1}{9}\right)$, т.е. вероятность того что случайная точка (X,Y) будет находиться в круге радиуса $R_1=1/3$ (см. рис. 5.5), можно было найти по формуле **(5.16)**:

$$P\left(X^2+Y^2<\frac{1}{9}\right)=\int_{-\frac{1}{3}}^{\frac{1}{3}}\int_{-\sqrt{\frac{1}{9}-x^2}}^{\sqrt{\frac{1}{9}-x^2}}\frac{1}{\pi}\,dx\,dy,$$

но проще это сделать, используя понятие «геометрической вероятности», т.е.

$$P\left(X^2+Y^2<\frac{1}{9}\right)=\left(\pi R_1^2\right)/\left(\pi R^2\right)=R_1^2/R^2=\left(\frac{1}{3}\right)^2/1^2=\frac{1}{9}\ .\blacktriangleright$$

5.4. Условные законы распределения. Числовые характеристики двумерной случайной величины. Регрессия

Определение. ***Условным законом распределения*** *одной из одномерных составляющих двумерной случайной величины* (X, Y) *называется ее закон распределения, вычисленный при условии, что другая составляющая приняла определенное значение (или попала в какой-то интервал).*

Выше (§ 5.1) рассмотрено нахождение условных распределений для дискретных случайных величин. Там же приведены формулы **(5.1)** и **(5.2)** условных вероятностей:

$$p_j(x_i)=\frac{P\left[(X=x_i)(Y=y_j)\right]}{P(Y=y_j)},\quad p_i(y_j)=\frac{P\left[(X=x_i)(Y=y_j)\right]}{P(X=x_i)}.$$

В случае непрерывных случайных величин необходимо определить *плотности вероятности* условных распределений $\varphi_y(x)$ и $\varphi_x(y)$. С этой целью в приведенных формулах заменим вероятности событий их «элементами вероятности», т.е. $P[(X=x_i)(Y=y_j)]$ на $\varphi(x,y)\,dx\,dy$, $P(X=x_i)$ на $\varphi(x)\,dx$, $P(Y=y_j)$ на $\varphi(y)\,dy$, $P_j(x_i)$ на $\varphi_y(x)\,dx$ и $P_i(y_j)$ на $\varphi_x(y)\,dy$, после сокращения на dx и dy получим:

$$\varphi_y(x)=\frac{\varphi(x,y)}{\varphi_2(y)},\quad \varphi_x(y)=\frac{\varphi(x,y)}{\varphi_1(x)}, \qquad (5.22)$$

т.е. *условная плотность вероятности*[1] *одной из одномерных составляющих двумерной случайной величины равна отношению ее совместной плотности к плотности вероятности другой составляющей.*

Соотношения **(5.22)**, записанные в виде

$$\varphi(x,y)=\varphi_1(x)\varphi_x(y)=\varphi_2(y)\varphi_y(x), \qquad (5.23)$$

называются *теоремой (правилом) умножения плотностей распределений.*

Используя формулы **(5.20)**, условные плотности вероятностей **(5.22)** можно выразить через совместную плотность следующим образом:

[1] Для условных плотностей вероятности используются также обозначения $\varphi(x \mid y)$, $\varphi(y \mid x)$.

$$\varphi_y(x)=\frac{\varphi(x,y)}{\int_{-\infty}^{+\infty}\varphi(x,y)\,dx},\quad \varphi_x(y)=\frac{\varphi(x,y)}{\int_{-\infty}^{+\infty}\varphi(x,y)\,dy}. \qquad (5.24)$$

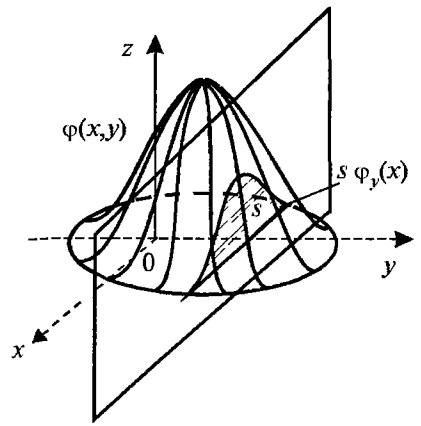


Рис. 5.9

Если, как отмечено выше, совместная плотность $\varphi(x,y)$ двумерной случайной величины представляет собой геометрически некоторую поверхность распределения, то, например, *условная плотность* $\varphi_y(x)$ *есть кривая распределения, подобная сечению этой поверхности плоскостью* $Y = y$, *параллельной координатной плоскости* Oxz *и отсекающей на оси* Oy *отрезок* y (рис. 5.9), и в соответствии с **(5.24)** и замечанием на с. 190 *получается из нее делением всех ординат на площадь данного сечения* s (т.е. сечение поверхности распределения есть кривая $s\varphi_y(x)$, где $0 \le s \le 1$) (рис. 5.9).

Аналогично можно пояснить геометрически и смысл условной плотности $\varphi_x(y)$.

Условные плотности $\varphi_y(x)$ и $\varphi_x(y)$ обладают всеми свойствами «безусловной» плотности, рассмотренной в гл. 3.

При изучении двумерных случайных величин рассматриваются числовые характеристики одномерных составляющих X и Y — математические ожидания и дисперсии. Для непрерывной случайной величины (X, Y) они определяются по формулам (получаемым из **(3.25)**, **(3.26)** с учетом **(5.20)**):

$$a_x=M(X)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}x\,\varphi(x,y)\,dx\,dy, \qquad (5.25)$$

$$a_y=M(Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}y\,\varphi(x,y)\,dx\,dy, \qquad (5.26)$$

$$D(X)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-a_x)^2\varphi(x,y)\,dx\,dy, \qquad (5.27)$$

$$D(Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(y-a_y)^2\varphi(x,y)\,dx\,dy. \qquad (5.28)$$

Наряду с ними рассматриваются также числовые характеристики условных распределений: *условные математические ожидания* $M_x(Y)$ и $M_y(X)$ и *условные дисперсии* $D_x(Y)$ и $D_y(X)$. Эти характеристики находятся по обычным формулам математического ожидания **(3.3)**, **(3.4)** **(3.25)** и дисперсии **(3.11)**, **(3.12)**, **(3.26)**, в которых вместо вероятностей событий p_i, p_j или плотностей вероятности $\varphi_1(x)$, $\varphi_2(y)$ используются условные вероятности $p_j(x_i)$, $p_i(y_j)$ или условные плотности вероятности $\varphi_y(x)$, $\varphi_x(y)$.

Например, для непрерывной случайной величины (X, Y)

$$M_x(Y)=\int_{-\infty}^{+\infty} y\,\varphi_x(y)dy, \qquad \textbf{(5.28′)}$$

$$D_x(Y)=\int_{-\infty}^{+\infty}[y-M_x(Y)]^2\varphi_x(y)dy.$$

Условное математическое ожидание случайной величины Y при $X = x$, т.е. $M_x(Y)$, есть функция от x, называемая *функцией регрессии* или просто *регрессией Y по X*; аналогично $M_y(X)$ называется *функцией регрессии* или просто *регрессией X по Y*. Графики этих функций называются соответственно *линиями регрессии* (или *кривыми регрессии*) *Y по X* и *X по Y*[1].

Основные свойства условных математических ожиданий и дисперсий аналогичны свойствам их «безусловных» аналогов, отмеченных выше (см. § 3.3, 3.4); при этом надо учитывать, что проводимые в них операции понимаются теперь уже не как действия над числами, а как действия над *функциями*.

Отметим здесь некоторые дополнительные с в о й с т в а условного математического ожидания:

1. *Если* $Z = g(X)$, *где* g — *некоторая неслучайная функция от* X, *то*
$$M_z(M_x(Y)) = M_z(Y).$$
В частности,
$$M(M_x(Y)) = M(Y)$$
(*правило повторного ожидания*).

2. *Если* $Z = g(X)$, *то* $M_x(ZY) = ZM_x(Y)$.

3. *Если случайные величины X и Y независимы, то*
$$M_x(Y) = M(Y).$$

□ Докажем в качестве примера (для непрерывной случайной величины) правило повторного ожидания $M(M_x(Y)) = M(Y)$. Условное математическое ожидание $M_x(Y)$ находится по условному

[1] Говорят также: *Y на X* и *X на Y*.

распределению случайной величины Y (при условии, что $X = x$), задаваемому условной плотностью вероятности $\varphi_x(Y)$. По формуле **(5.28′)** с учетом **(5.22)** получим

$$M_x(Y)=\int_{-\infty}^{+\infty} y\,\varphi_x(y)dy=\int_{-\infty}^{+\infty} y\frac{\varphi(x,\,y)}{\varphi_1(x)}dy.$$

Теперь

$$M(M_x(Y))=\int_{-\infty}^{+\infty}M_x(Y)\varphi_1(x)dx=\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} y\frac{\varphi(x,\,y)}{\varphi_1(x)}dy\right)\varphi_1(x)dx=$$

$$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y\varphi(x,y)dx\,dy=M(Y)$$

(в соответствии с формулой **(5.26)**). ■

5.5. Зависимые и независимые случайные величины

Выше (§ 3.2) введено понятие независимости дискретных случайных величин X и Y, основанное на независимости связанных с ними событий $X = x_i$ и $Y = y_j$ при любых i и j. Теперь можно дать общее определение независимости случайных величин, основанное на независимости событий $X<x$ и $Y<y$, т.е. функций распределений $F_1(x)$ и $F_2(y)$.

О п р е д е л е н и е. *Случайные величины X и Y называются* ***независимыми****, если их совместная функция распределения* $F(x,y)$ *представляется в виде произведения функций распределений* $F_1(x)$ *и* $F_2(y)$ *этих случайных величин*, т.е.

$$F(x,y)=F_1(x)\cdot F_2(y). \qquad \textbf{(5.29)}$$

В противном случае, при невыполнении равенства **(5.29)**, случайные величины X и Y называются ***зависимыми***.

Дифференцируя дважды равенство **(5.29)** по аргументам x и y, получим

$$\varphi(x,y)=\varphi_1(x)\cdot\varphi_2(y), \qquad \textbf{(5.30)}$$

т.е. *для независимых непрерывных случайных величин X и Y их совместная плотность* $\varphi(x,y)$ *равна произведению плотностей вероятности* $\varphi_1(x)$ *и* $\varphi_2(y)$ *этих случайных величин.*

Сравнивая формулы **(5.30)** и **(5.23)**, можно утверждать, что *независимость двух случайных величин X и Y означает, что условные плотности вероятности каждой из них совпадают с соответствующими «безусловными» плотностями*, т.е.

$$\varphi_y(x)=\varphi_1(x) \text{ и } \varphi_x(y)=\varphi_2(y). \qquad \textbf{(5.31)}$$

До сих пор мы сталкивались с понятием функциональной зависимости между переменными X и Y, когда каждому значению x одной переменной соответствовало строго определенное значение y другой. Например, зависимость между двумя случайными величинами — числом вышедших из строя единиц оборудования за определенный период времени и их стоимостью — функциональная.

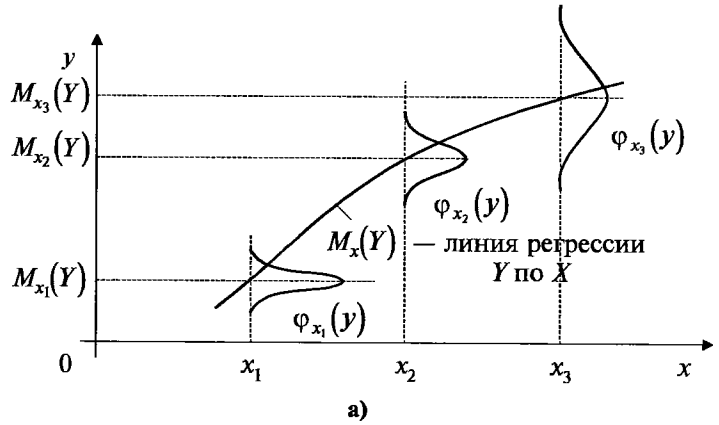
В общем случае, при невыполнении условий **(5.29)—(5.31)**, сталкиваются с зависимостью другого типа, менее жесткой, чем функциональная.

Определение. *Зависимость между двумя случайными величинами называется* ***вероятностной*** *(****стохастической*** *или* ***статистической****), если каждому значению одной из них соответствует определенное (условное) распределение другой.*

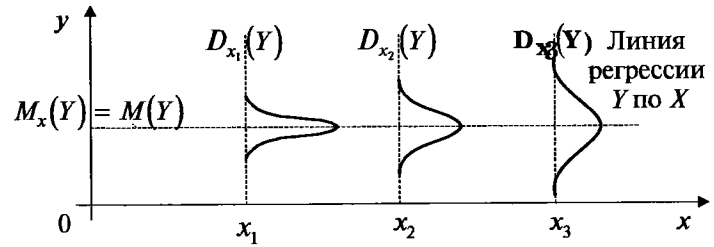
В случае вероятностной (стохастической) зависимости нельзя, зная значение одной из них, точно определить значение другой, а можно указать лишь распределение другой величины. Например, зависимости между числом отказов оборудования и затрат на его профилактический ремонт, весом и ростом человека, затратами времени школьника на просмотр телевизионных передач и чтение книг и т.п. являются вероятностными (стохастическими).

На рис. 5.10 приведены примеры зависимых и независимых случайных величин X и Y.

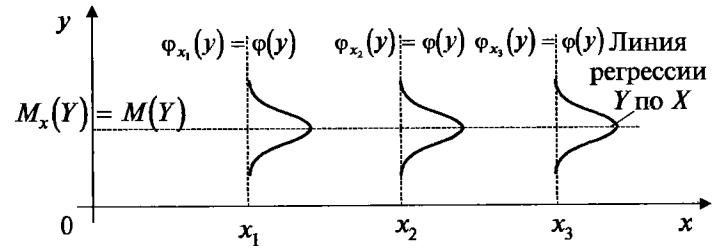
На рис. 5.10*а* зависимость между X и Y проявляется в том, что с изменением x меняется как распределение Y, так и условное математическое ожидание $M_x(Y)$ (с увеличением x $M_x(Y)$ увеличивается). Там же показана зависимость $M_x(Y)$ от x, т.е. линия регрессии Y по X.



а)



б)



в)

Рис. 5.10

На рис. 5.10*б* зависимость между случайными величинами проявляется в изменении условных дисперсий (с ростом x $D_x(Y)$ увеличивается), при этом $M_x(Y) = \text{const}$, т.е. линия регрессии Y по X параллельна оси Ox. На рис. 5.10*в* случайные величины Y и X независимы, так как с изменением x распределение случайной величины Y, а значит, условное математическое ожидание $M_x(Y)$ и условная дисперсия $D_x(Y)$ не меняются.

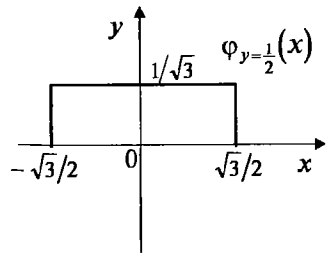


Рис. 5.11

Таким образом, обобщая, можно утверждать, что *если случайные величины Y и X независимы, то линии регрессии Y по X и X по Y параллельны координатным осям Ox и Oy.*

▷ **Пример 5.4.** По данным примера **5.3** определить: а) условные плотности случайных величин X и Y; б) зависимы или независимы случайные величины X и Y; в) условные математические ожидания и условные дисперсии.

Решение. а) Найдем условную плотность $\varphi_y(x)$ по формуле **(5.22)**, учитывая, что $\varphi_2(y) \neq 0$.

$$\varphi_y(x)=\frac{\varphi(x,y)}{\varphi_2(y)}=\begin{cases}\dfrac{1}{2\sqrt{1-y^2}} & \text{при } |x|<\sqrt{1-y^2},\\ 0 & \text{при } |x|>\sqrt{1-y^2}.\end{cases}$$

График $\varphi_y(x)$ при $y=1/2$ показан на рис. 5.11.

Аналогично

$$\varphi_x(y)=\frac{\varphi(x,y)}{\varphi_1(x)}=\begin{cases}\dfrac{1}{2\sqrt{1-x^2}} & \text{при } |y|<\sqrt{1-x^2},\\ 0 & \text{при } |y|>\sqrt{1-x^2}.\end{cases}$$

б) X и Y — зависимые случайные величины, так как $\varphi(x,y)\neq\varphi_1(x)\varphi_2(y)$ или $\varphi_y(x)\neq\varphi_1(x)$, $\varphi_x(y)\neq\varphi_2(y)$ (см. пример **5.3** и п. а)).

в) Найдем условное математическое ожидание $M_x(Y)$, учитывая, что $\varphi_x(y)=\dfrac{1}{2\sqrt{1-x^2}}$.

$$M_x(Y)=\int_{-\infty}^{+\infty} y\,\varphi_x(y)\,dy=\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}\frac{1}{2\sqrt{1-x^2}}\,y\,dy=$$

$$=\frac{1}{2\sqrt{1-x^2}}\cdot\frac{y^2}{2}\Bigg|_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}=0.$$

Аналогично $M_y(X)=\int_{-\infty}^{+\infty} x\,\varphi_y(x)\,dx=0$.

Этот результат очевиден в силу того, что круг $x^2+y^2\le 1$ (рис. 5.5) симметричен относительно координатных осей. Таким образом, линия регрессии Y по X совпадает с осью Ox ($M_x(Y)=0$), а линия регрессии X по Y — с осью Oy ($M_y(X)=0$).

Найдем условную дисперсию $D_x(Y)$:

$$D_x(Y)=\int_{-\infty}^{+\infty}\left[y-M_x(Y)\right]^2\varphi_x(y)\,dy=\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}\frac{1}{2\sqrt{1-x^2}}\,y^2\,dy=$$

$$=\frac{1}{2\sqrt{1-x^2}}\,\frac{y^3}{3}\Bigg|_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}=\frac{1}{2\sqrt{1-x^2}}\cdot\frac{1}{3}\cdot 2\sqrt{\left(1-x^2\right)^3}=\frac{1-x^2}{3}\ (0\le x\le 1).$$

(Тот же результат можно получить проще — по формуле дисперсии равномерного закона распределения (**4.20**):

$$D_x(Y)=\frac{(b-a)^2}{12}=\frac{\left[\sqrt{1-x^2}-(-\sqrt{1-x^2})\right]^2}{12}=\frac{1-x^2}{3})$$

Аналогично

$$D_y(X)=\frac{1-y^2}{3}\ (0\le y\le 1).$$

Таким образом, по мере удаления от начала координат дисперсия условных распределений уменьшается от 1/3 до 0. ▶

5.6. Ковариация и коэффициент корреляции

Пусть имеется двумерная случайная величина (X,Y), распределение которой известно, т.е. известна табл. 5.1 или совместная плотность вероятности $\varphi(x,y)$. Тогда можно найти (см. § 5.4) математические ожидания $M(X)=a_x$, $M(Y)=a_y$ и дисперсии $D(X)=\sigma_x^2$ и $D(Y)=\sigma_y^2$ одномерных составляющих X и Y. Однако математические ожидания и дисперсии случайных величин X и Y недостаточно полно характеризуют двумерную случайную величину (X,Y), так как не выражают степени зависимости ее составляющих X и Y. Эту роль выполняют **ковариация** и **коэффициент корреляции.**

О п р е д е л е н и е. ***Ковариацией*** (или ***корреляционным моментом***) K_{xy} *случайных величин* X *и* Y *называется математическое ожидание произведения отклонений этих величин от своих математических ожиданий*[1], т.е.

$$K_{xy}=M[(X-M(X))(Y-M(Y))], \qquad \textbf{(5.32)}$$

или $$K_{xy}=M[(X-a_x)(Y-a_y)],$$

где $a_x=M(X)$, $a_y=M(Y)$.

Из определения следует, что $K_{xy}=K_{yx}$. Кроме того,

$$K_{xx}=M[(X-a_x)(X-a_x)]=M(X-a_x)^2=D(X),$$

т.е. *ковариация случайной величины с самой собой есть ее дисперсия.*

[1] Ковариацию называют еще *вторым смешанным центральным моментом* случайных величин X и Y. Для ковариации X и Y используются также обозначения $\mathrm{cov}(X,Y)$, $\sigma_{(xy)}$.

Для дискретных случайных величин

$$K_{xy} = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i - a_x)(y_j - a_y)p_{ij}. \qquad (5.33)$$

Для непрерывных случайных величин

$$K_{xy} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x - a_x)(y - a_y)\varphi(x,y)\,dx\,dy. \qquad (5.34)$$

Ковариация двух случайных величин характеризует как степень зависимости случайных величин, так и их рассеяние вокруг точки (a_x, a_y). Об этом, в частности, свидетельствуют свойства ковариации случайных величин.

1. *Ковариация двух независимых случайных величин равна нулю.*

□ Для независимых случайных величин согласно **(5.26)** $\varphi(x,y) = \varphi_1(x)\varphi_2(y)$. Поэтому формула ковариации, например, **(5.34)** для непрерывных случайных величин примет вид:

$$K_{xy} = \int_{-\infty}^{+\infty}(x - a_x)\varphi_1(x)\,dx\int_{-\infty}^{+\infty}(y - a_y)\varphi_2(y)\,dy = 0,$$

так как каждый из полученных интегралов есть центральный момент первого порядка, равный нулю. ■

2. *Ковариация двух случайных величин равна математическому ожиданию их произведения минус произведение математических ожиданий*, т.е.

$$K_{xy} = M(XY) - M(X)\cdot M(Y), \qquad (5.35)$$

или

$$K_{xy} = M(XY) - a_x a_y.$$

□ По определению **(5.32)**:

$$K_{xy} = M[(X - a_x)(Y - a_y)] = M(XY - a_x Y - a_y X + a_x a_y).$$

Учитывая, что математические ожидания $M(X) = a_x$ и $M(Y) = a_y$ — неслучайные величины, получим

$$K_{xy} = M(XY) - a_x M(Y) - a_y M(X) + a_x a_y =$$

$$= M(XY) - a_x a_y - a_y a_x + a_x a_y = M(XY) - a_x a_y. \blacksquare$$

3. *Ковариация двух случайных величин по абсолютной величине не превосходит произведения их средних квадратических отклонений*, т.е.

$$|K_{xy}| \le \sigma_x \sigma_y. \qquad (5.36)$$

□ Возьмем очевидное неравенство:

$$M\left(\frac{X - a_x}{\sigma_x} \pm \frac{Y - a_y}{\sigma_y}\right)^2 \ge 0 \text{ или}$$

$$M\left[\left(\frac{X - a_x}{\sigma_x}\right)^2 \pm 2\left(\frac{X - a_x}{\sigma_x}\cdot\frac{Y - a_y}{\sigma_y}\right) + \left(\frac{Y - a_y}{\sigma_y}\right)^2\right] =$$

$$= \frac{1}{\sigma_x^2}M(X - a_x)^2 \pm \frac{2}{\sigma_x\sigma_y}M[(X - a_x)(Y - a_y)] + \frac{1}{\sigma_y^2}M(Y - a_y)^2 =$$

$$= \frac{D(X)}{\sigma_x^2} \pm \frac{2K_{xy}}{\sigma_x\sigma_y} + \frac{D(Y)}{\sigma_y^2} = 2 \pm \frac{2K_{xy}}{\sigma_x\sigma_y} \ge 0 \qquad (5.37)$$

(учтено, что σ_x и σ_y — неслучайные величины и дисперсии

$$D(X) = M(X - a_x)^2 = \sigma_x^2,\ D(Y) = M(Y - a_y)^2 = \sigma_y^2).$$

Из неравенства **(5.37)** следует доказываемое **(5.36)**. ■

Ковариация, как уже отмечено, характеризует не только степень зависимости двух случайных величин, но и их разброс, рассеяние. Кроме того, она — величина размерная, ее размерность определяется произведением размерностей случайных величин. Это затрудняет использование ковариации для оценки степени зависимости для различных случайных величин. Этих недостатков лишен **коэффициент корреляции.**

Определение. ***Коэффициентом корреляции***[1] *двух случайных величин называется отношение их ковариации к произведению средних квадратических отклонений этих величин:*

$$\rho_{xy} = \frac{K_{xy}}{\sigma_x\sigma_y}. \qquad (5.38)$$

Из определения следует, что $\rho_{xy} = \rho_{yx} = \rho$. Очевидно также, что коэффициент корреляции есть *безразмерная* величина.

[1] Для коэффициента корреляции случайных величин в литературе используется также обозначение $\mathrm{corr}(X, Y)$.

Отметим с в о й с т в а коэффициента корреляции.

1. *Коэффициент корреляции принимает значения на отрезке* $[-1;1]$, т.е.

$$-1 \le \rho \le 1. \qquad (5.39)$$

□ Из неравенства **(5.37)**:

$$2 \pm \frac{2K_{xy}}{\sigma_x \sigma_y} = 2 \pm 2\rho \ge 0, \text{ откуда } -1 \le \rho \le 1. \blacksquare$$

2. *Если случайные величины независимы, то их коэффициент корреляции равен нулю*, т.е. $\rho = 0$.

□ $\rho = 0$, так как в этом случае $K_{xy} = 0$. ■

Случайные величины называются *некоррелированными*, если их коэффициент корреляции равен нулю. Таким образом, из *независимости случайных величин следует их некоррелированность.*

Обратное утверждение, вообще говоря, неверно: из *некоррелированности двух случайных величин еще не следует их независимость* (см. пример **5.6**).

3. *Если коэффициент корреляции двух случайных величин равен (по абсолютной величине) единице, то между этими случайными величинами существует линейная функциональная зависимость.*

□ Выше было получено, что

$$M\left(\frac{X - a_x}{\sigma_x} \pm \frac{Y - a_y}{\sigma_y}\right)^2 = 2 \pm \frac{2K_{xy}}{\sigma_x \sigma_y} = 2 \pm 2\rho.$$

Если $\rho = \mp 1$, то $2 \pm 2\rho = 0$ и $M\left(\frac{X - a_x}{\sigma_x} \pm \frac{Y - a_y}{\sigma_y}\right)^2 = 0.$

Равенство математического ожидания неотрицательной случайной величины нулю означает, что сама случайная величина тождественно равна нулю.

$$\frac{X - a_x}{\sigma_x} \pm \frac{Y - a_y}{\sigma_y} = 0 \text{ при } \rho = \mp 1 \text{ или}$$

$$Y = a_y + \frac{\sigma_y}{\sigma_x}(X - a_x) \text{ при } \rho = 1 \text{ и } Y = a_y - \frac{\sigma_y}{\sigma_x}(X - a_x)$$

при $\rho = -1$, т.е. X и Y связаны линейной ф у н к ц и о н а л ь н о й зависимостью. ■

▷ **Пример 5.5.** По данным примера **5.2** определить ковариацию и коэффициент корреляции случайных величин X и Y.

Р е ш е н и е. В примере **5.2** были получены следующие законы распределения одномерных случайных величин:

X:

x_i	1	2
p_i	0,8	0,2

и

Y:

y_j	−1	0	1	2
p_j	0,2	0,3	0,3	0,2

Найдем математические ожидания и средние квадратические отклонения этих случайных величин:

$$a_x = M(X) = \sum_{i=1}^{2} x_i p_i = 1 \cdot 0{,}8 + 2 \cdot 0{,}2 = 1{,}2;$$

$$M(X^2) = \sum_{i=1}^{2} x_i^2 p_i = 1^2 \cdot 0{,}8 + 2^2 \cdot 0{,}2 = 1{,}6;$$

$$D(X) = M(X^2) - a_x^2 = 1{,}6 - 1{,}2^2 = 0{,}16; \quad \sigma_x = \sqrt{D(X)} = \sqrt{0{,}16} = 0{,}4;$$

$$a_y = M(Y) = \sum_{j=1}^{4} y_j p_j = (-1) \cdot 0{,}2 + 0 \cdot 0{,}3 + 1 \cdot 0{,}3 + 2 \cdot 0{,}2 = 0{,}5;$$

$$M(Y^2) = \sum_{j=1}^{4} y_j^2 p_j = (-1)^2 \cdot 0{,}2 + 0^2 \cdot 0{,}3 + 1^2 \cdot 0{,}3 + 2^2 \cdot 0{,}2 = 1{,}3;$$

$$D(Y) = M(Y^2) - a_y^2 = 1{,}3 - 0{,}5^2 = 1{,}05; \quad \sigma_y = \sqrt{D(Y)} = \sqrt{1{,}05} = 1{,}025.$$

Для нахождения математического ожидания $M(XY)$ произведения случайных величин X и Y можно было составить закон распределения произведения двух дискретных случайных величин (с вероятностями его значений из табл. 5.2)[1], а затем по нему найти $M(XY)$. Но делать это вовсе не обязательно. $M(XY)$ можно найти непосредственно по табл. 5.2 распределения двумерной случайной величины (X, Y) по формуле:

[1] Закон распределения (XY) имеет вид:

$(xy)_k$	−2	−1	0	1	2	4
p_k	0,1	0,1	0,3	0,3	0,15	0,05

$$M(XY)=\sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j p_{ij},$$

где двойная сумма $\sum_{i=1}^{n}\sum_{j=1}^{m}$ означает суммирование по всем nm клеткам таблицы (n — число строк, m — число столбцов):

$$M(XY) = 1\cdot(-1)\cdot 0{,}10+(-1)\cdot 0\cdot 0{,}25+1\cdot 1\cdot 0{,}30+1\cdot 2\cdot 0{,}15+$$
$$+2\cdot(-1)\cdot 0{,}10+2\cdot 0\cdot 0{,}05+2\cdot 1\cdot 0+2\cdot 2\cdot 0{,}05 = 0{,}5.$$

Вычислим ковариацию K_{xy} по формуле **(5.35)**:

$$K_{xy} = M(XY) - a_x a_y = 0{,}5 - 1{,}2\cdot 0{,}5 = -0{,}1.$$

Вычислим коэффициент корреляции ρ по формуле **(5.38)**:

$$\rho = \frac{K_{xy}}{\sigma_x \sigma_y} = \frac{-0{,}1}{0{,}4\cdot 1{,}025} = -0{,}244,$$

т.е. между случайными величинами X и Y существует отрицательная линейная зависимость; следовательно, при увеличении (уменьшении) одной из случайных величин другая имеет некоторую тенденцию уменьшаться (увеличиваться). ▶

▷ **Пример 5.6.** По данным примера **5.3** определить: а) ковариацию и коэффициент корреляции случайных величин X и Y; б) коррелированы или некоррелированы эти случайные величины.

Р е ш е н и е. а) Вначале по формулам **(5.25)**, **(5.26)** найдем математические ожидания $a_x = M(X)$ и $a_y = M(Y)$.

$$a_x = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x\varphi(x,y)\,dx\,dy = \frac{1}{\pi}\int_{-1}^{1} x\,dx\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} dy = 0.$$

Аналогично $a_y = 0$ (то, что $a_x = a_y = 0$, очевидно из соображения симметрии распределения в круге, из которой следует, что центр его массы лежит в начале координат).

По формуле **(5.34)** ковариация

$$K_{xy} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-a_x)(y-a_y)\varphi(x,y)\,dx\,dy =$$
$$= \int_{-1}^{1}\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} xy\,\frac{1}{\pi}\,dx\,dy = \frac{1}{\pi}\int_{-1}^{1} x\,dx\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} y\,dy =$$
$$= \frac{1}{\pi}\int_{-1}^{1} x\left(\frac{y^2}{2}\Bigg|_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}\right)dx = 0.$$

Соответственно коэффициент корреляции $\rho = \frac{K_{xy}}{\sigma_x\sigma_y} = 0.$

б) Так как $\rho = 0$, то случайные величины X и Y некоррелированы. В примере **5.4** установлено, что эти случайные величины зависимы; таким образом, наглядно убеждаемся в том, что из некоррелированности величин еще не вытекает их независимость. ▶

С помощью ковариации можно дополнить и уточнить некоторые **свойства математического ожидания и дисперсии** (рассмотренные в гл. 3).

1. *Математическое ожидание произведения двух случайных величин равно сумме произведения их математических ожиданий и ковариации этих случайных величин*:

$$M(XY)=M(X)\cdot M(Y)+K_{xy}. \qquad \textbf{(5.40)}$$

□ Формула **(5.40)** следует непосредственно из формулы **(5.35)**. ■

Если $K_{xy} = 0$, то

$$M(XY) = M(X)M(Y), \qquad \textbf{(5.41)}$$

т.е. *математическое ожидание произведения двух некоррелированных случайных величин равно произведению их математических ожиданий*[1].

2. *Дисперсия суммы двух случайных величин равна сумме их дисперсий плюс удвоенная ковариация этих случайных величин*:

$$D(X+Y) = D(X)+D(Y)+2K_{xy}. \qquad \textbf{(5.42)}$$

□ Пусть $Z = X+Y$. По свойству математического ожидания $a_z = a_x+a_y$. Поэтому $Z-a_z = (X-a_x)+(Y-a_y)$.

По определению дисперсии

$$D(X+Y)=D(Z)=M(Z-a_z)^2=M(X-a_x)^2+2M[(X-a_x)(Y-a_y)]+$$
$$+M(Y-a_y)^2=D(X)+2K_{xy}+D(Y). \blacksquare$$

Формула **(5.42)** может быть обобщена на любое число слагаемых:

$$D\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} D(X_i) + 2\sum_{i=1}^{n}\sum_{j=1}^{m} K_{ij}, \qquad \textbf{(5.43)}$$

где K_{ij} — ковариации случайных величин X_i и X_j.

[1] В § 3.3 свойство **(5.41)** сформулировано для независимых случайных величин. Теперь выясняется, что в случае *двух* сомножителей достаточно менее жесткого требования некоррелированности случайных величин. В случае *произвольного числа* сомножителей, как показывает анализ, требование независимости случайных величин должно быть сохранено.

Для некоррелированных (и, разумеется, для независимых) случайных величин

$$D\left(\sum_{i=1}^{n} X_i\right)=\sum_{i=1}^{n} D(X_i), \qquad (5.44)$$

т.е. *дисперсия суммы некоррелированных случайных величин равна сумме их дисперсий.*

5.7. Двумерный (n-мерный) нормальный закон распределения

Определение. *Случайная величина (X,Y) называется распределенной по **двумерному нормальному закону**, если ее совместная плотность имеет вид:*

$$\varphi_N(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-L(x,y)}, \qquad (5.45)$$

где

$$L(x,y)=\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-a_x}{\sigma_x}\right)^2-2\rho\frac{x-a_x}{\sigma_x}\cdot\frac{y-a_y}{\sigma_y}+\left(\frac{y-a_y}{\sigma_y}\right)^2\right]. \qquad (5.46)$$

Из определения (5.45), (5.46) следует, что *двумерный нормальный закон распределения определяется пятью параметрами:* a_x, a_y, σ_x^2, σ_y^2, ρ.

Для выяснения теоретико-вероятностного смысла этих параметров по формулам (5.25)—(5.28), (5.34) найдем математические ожидания случайных величин $M(X)$ и $M(Y)$, их дисперсии $D(X)$ и $D(Y)$ и ковариацию K_{xy}. Получим после замены $\varphi_N(x,y)$ его выражением (5.45) с учетом (5.46) и вычисления интегралов[1]:

$$M(X)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}x\varphi_N(x,y)\,dx\,dy=a_x$$

и аналогично $M(Y)=a_y$;

$$D(X)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-a_x)^2\varphi_N(x,y)\,dx\,dy=\sigma_x^2,$$

[1] Вычисление соответствующих несобственных интегралов в § 5.7 опускаем.

аналогично $D(Y)=\sigma_y^2$;

$$K_{xy}=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-a_x)(y-a_y)\varphi_N(x,y)\,dx\,dy=\rho\sigma_x\sigma_y.$$

Таким образом, параметры a_x и a_y выражают математические ожидания случайных величин X и Y, параметры σ_x^2 и σ_y^2 — их дисперсии, а $\frac{K_{xy}}{\sigma_x\sigma_y}=\rho$ — коэффициент корреляции между случайными величинами X и Y.

Найдем плотности вероятности одномерных случайных величин X и Y по формулам (5.20) с учетом (5.45), (5.46):

$$\varphi_1(x)=\int_{-\infty}^{+\infty}\varphi_N(x,y)\,dy=\frac{1}{\sigma_x\sqrt{2\pi}}e^{-\frac{(x-a_x)^2}{2\sigma_x^2}} \qquad (5.47)$$

и аналогично

$$\varphi_2(y)=\frac{1}{\sigma_y\sqrt{2\pi}}e^{-\frac{(y-a_y)^2}{2\sigma_y^2}} \qquad (5.48)$$

Как и следовало ожидать, каждый из законов распределения одномерных случайных величин X и Y является нормальным с параметрами соответственно (a_x,σ_x^2) и (a_y,σ_y^2).

Найдем условные плотности вероятности случайных величин X и Y по формулам (5.22) с учетом (5.45)—(5.48):

$$\varphi_y(x)=\frac{\varphi_N(x,y)}{\varphi_2(y)}=\frac{1}{\sigma_x\sqrt{1-\rho^2}\sqrt{2\pi}}e^{-\frac{1}{2(1-\rho^2)}\left(\frac{x-a_x}{\sigma_x}-\rho\frac{y-a_y}{\sigma_y}\right)^2}$$

и аналогично $\varphi_x(y)=\dfrac{1}{\sigma_y\sqrt{1-\rho^2}\sqrt{2\pi}}e^{-\frac{1}{2(1-\rho^2)}\left(\frac{y-a_y}{\sigma_y}-\rho\frac{x-a_x}{\sigma_x}\right)^2}$

Нетрудно убедиться в том, что каждый из условных законов распределения случайных величин X и Y является нормальным с условным математическим ожиданием и условной дисперсией, определяемыми по формулам:

$$M_y(X)=a_x+\rho\frac{\sigma_x}{\sigma_y}(y-a_y), \qquad (5.49)$$

$$D_y(X)=\sigma_x^2(1-\rho^2), \qquad (5.50)$$

$$M_x(Y)=a_y+\rho\frac{\sigma_y}{\sigma_x}(x-a_x), \qquad (5.51)$$

$$D_x(Y)=\sigma_y^2(1-\rho^2). \qquad (5.52)$$

Из формул **(5.49)**, **(5.51)** следует, что *линии регрессии* $M_y(X)$ и $M_x(Y)$ *нормально распределенных случайных величин представляют собой прямые линии, т.е. нормальные регрессии* Y *по* X *и* X *по* Y *всегда линейны.*

Из **(5.50)** и **(5.52)** следует, что условные дисперсии $D_y(X)$ и $D_x(Y)$ (а значит, и условные стандартные отклонения $\sigma_y(X)$ и $\sigma_x(Y)$) постоянны и не зависят от значений y или x. Это свойство называется *гомоскедастичностью* или *равноизменчивостью* условных нормальных распределений и имеет существенное значение в статистическом анализе.

Теорема. *Если две нормально распределенные случайные величины* X *и* Y *некоррелированы, то они независимы.*

□ По условию коэффициент корреляции $\rho = 0$. Найдем выражение совместной плотности. По формулам **(5.45)**, **(5.46)** при $\rho = 0$ получим:

$$\varphi_N(x,y)=\frac{1}{\sigma_x\sqrt{2\pi}}e^{\frac{(x-a_x)^2}{2\sigma_x^2}}\cdot\frac{1}{\sigma_y\sqrt{2\pi}}e^{\frac{(y-a_y)^2}{2\sigma_y^2}}=\varphi_1(x)\varphi_2(y), \qquad (5.53)$$

где $\varphi_1(x)$ и $\varphi_2(y)$ — плотности вероятностей одномерных случайных величин X и Y.

В соответствии с условием **(5.30)** это означает независимость случайных величин X и Y. ■

Таким образом, *для нормально распределенных случайных величин термины «некоррелированность» и «независимость» равносильны.*

Поверхность $\varphi_N(x,y)$ нормального распределения представляет собой холмообразную поверхность, называемую иногда «палаткой Гаусса» (рис. 5.12).

Сечения этой поверхности плоскостями $x = a$ (перпендикулярно оси Ox) и $y = b$ (перпендикулярно оси Oy) имеют форму нормальных кривых с центрами, лежащими на линии регрессии Y по X **(5.51)** и X по Y **(5.53)**, и со средними квадратическими отклонениями, равными $\sigma_x\sqrt{1-\rho^2}$ и $\sigma_y\sqrt{1-\rho^2}$.

Сечение поверхности нормального распределения плоскостью $z = c$ (где $0 < c < \varphi_N(a_x,a_y)$), параллельной плоскости Oxy, представляет эллипс, называемый *эллипсом рассеяния* (рис. 5.13):

$$\frac{(x-a_x)^2}{\sigma_x^2}-2\rho\frac{(x-a_x)(y-a_y)}{\sigma_x\sigma_y}+\frac{(y-a_y)^2}{\sigma_y^2}=a^2,$$

где $a^2=-2(1-\rho^2)\ln(2\pi c\,\sigma_x\sigma_y\sqrt{1-\rho^2})$.

(В силу ограничения для c аргумент логарифма меньше 1, а само значение логарифма отрицательно.)

Центр эллипса находится в точке (a_x, a_y), а его оси образуют с осью Ox углы α и $\pi/2+\alpha$, где α определяется из условия

$$\operatorname{tg}2\alpha=\frac{2\rho\,\sigma_x\sigma_y}{\sigma_x^2-\sigma_y^2}.$$

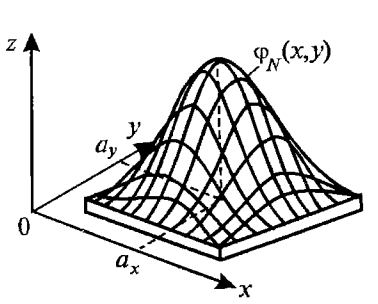


Рис. 5.12

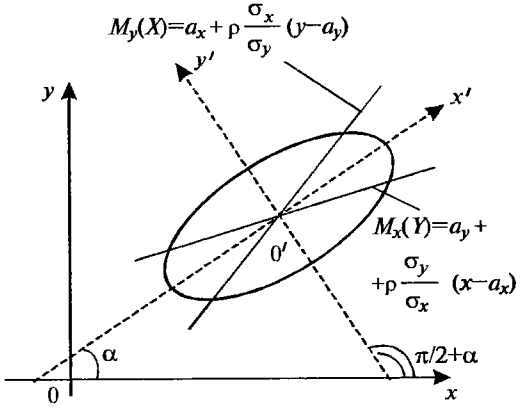


Рис. 5.13

О п р е д е л е н и е. *Случайная величина* $(X_1,X_2,...,X_n)$ *называется распределенной по* ***n*-мерному нормальному закону**, *если ее совместная плотность имеет вид*:

$$\varphi_N(x_1,x_2,...,x_n)=\frac{1}{\left(2\pi^{n/2}\sqrt{|K_{ij}|}\right)}e^{-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_{ij}^{(-1)}(x_i-a_i)(x_j-a_j)}, \qquad (5.54)$$

где a_i — математическое ожидание одномерной составляющей X_i $(i = 1,2,...,n)$;

$|K_{ij}|$ — определитель ковариационной матрицы (K_{ij}) случайной величины $(X_1, X_2, ..., X_n)$:

$$(K_{ij}) = \begin{pmatrix} K_{11} & K_{12} & \cdots & K_{1n} \\ K_{21} & K_{22} & \cdots & K_{2n} \\ - & - & - & - \\ K_{n1} & K_{n2} & \cdots & K_{nn} \end{pmatrix}, \qquad (5.55)$$

K_{ij}^{-1} — элементы матрицы, обратной по отношению к ковариационной матрице (K_{ij}).

Из определения **(5.54)**, **(5.55)** следует, что нормальный закон распределения n-мерной случайной величины (n-мерного случайного вектора) $X = (X_1, X_2, ..., X_n)$ характеризуется параметрами, задаваемыми вектором математических ожиданий $a = (a_1, a_2, ..., a_n)'$ и ковариационной матрицей (K_{ij}), где $K_{ij} = M[X_i - a_i)(X_j - a_j)]$ $i, j = 1,2, ..., n$.

Ковариационная матрица и ее определитель, называемый *обобщенной дисперсией n-мерной случайной величины*, являются аналогами дисперсии одномерной случайной величины и характеризуют степень *случайного разброса отдельно по каждой составляющей и в целом по n-мерной величине.*

Можно показать, что при $n = 2$ совместная плотность **(5.54)** имеет вид **(5.45)**, **(5.46)** — совместной плотности двумерного нормального распределения, а свойства n-мерного нормального закона аналогичны свойствам двумерного.

5.8. Функция случайных величин. Композиция законов распределения

Одной из важных задач в теории вероятностей является определение **закона распределения** функции одной или нескольких случайных величин, если известны распределения одного или нескольких аргументов. Построение законов распределения функций некоторых дискретных случайных величин $(kX, X^m, X+Y, X-Y, XY)$ рассмотрено в § 3.2. Обратимся теперь к непрерывным случайным величинам.

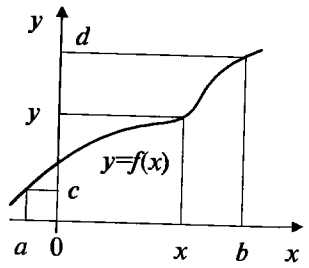


Рис. 5.14

Пусть имеется непрерывная случайная величина X с плотностью вероятности $\varphi(x)$, а случайная величина Y есть функция от X, т.е. $Y = f(X)$. Требуется найти закон распределения случайной величины Y.

Пусть функция $f(X)$ — строго монотонна, непрерывна и дифференцируема на отрезке $[a,b]$ и $f(a) = c$, $f(b) = d$ (рис. 5.14). Полагаем, что $f'(x) > 0$. Тогда функция распределения $G(y)$ случайной величины $Y = f(X)$:

$$G(y) = P(Y < y) = \begin{cases} 0 & \text{при } y < c, \\ \int_c^y g(y)\,dy & \text{при } c \le y \le d, \end{cases}$$

где $g(y)$ — плотность вероятности случайной величины $Y = f(X)$.

Если $c \le y \le d$,

$$G(y) = P(Y < y) = P(f(X) < y).$$

Неравенство $f(X) < y$ равносильно неравенству $X < f^{-1}(y)$, где $f^{-1}(y)$ — функция, обратная функции $f(x)$ на отрезке $[a,b]$. Поэтому

$$G(y) = P[X < f^{-1}(y)] = \int_d^{f^{-1}(y)} \varphi(x)\,dx.$$

По теореме о производной интеграла по переменному верхнему пределу[1]:

$$g(y) = G'(y) = \varphi[f^{-1}(y)] \left| [f^{-1}(y)]' \right|. \qquad (5.56)$$

(Производную $[f^{-1}(y)]'$ берем по абсолютной величине, так как в случае, если функция $f(x)$ на отрезке $[a,b]$ убывающая, то обратная ей функция $f^{-1}(y)$ убывающая и производная $[f^{-1}(y)]' < 0$, а плотность вероятности $g(y)$ отрицательной быть не может.

[1] Формула **(5.56)** остается в силе на бесконечном интервале $(-\infty, +\infty)$, т.е. при $a = -\infty, b = +\infty$.

Можно показать, что для нахождения **числовых характеристик** случайной величины $Y = f(X)$ не обязательно знать закон ее распределения, достаточно знать закон распределения аргумента:

$$a_y = M(Y) = M[f(X)] = \int_{-\infty}^{+\infty} f(x)\varphi(x)\,dx, \qquad \textbf{(5.57)}$$

$$D(Y) = D[f(X)] = \int_{-\infty}^{+\infty} [f(x) - a_y]^2 \varphi(x)\,dx. \qquad \textbf{(5.58)}$$

▷ **Пример 5.7.** Найти плотность вероятности случайной величины $Y = 1 - X^3$, где случайная величина X распределена по закону Коши с плотностью вероятности $\varphi(x) = \dfrac{1}{\pi(1+x^2)}$.

Р е ш е н и е. По условию $y = f(x) = 1 - x^3$, откуда $x = f^{-1}(y) = \sqrt[3]{1-y}$. Производная (по абсолютной величине)

$$\left|[f^{-1}(y)]'\right| = \frac{1}{3\sqrt[3]{(1-y)^2}}.$$

По формуле **(5.56)** плотность вероятности

$$g(y) = \frac{1}{3\pi\left(1+\sqrt[3]{(1-y)^2}\right)\sqrt[3]{(1-y)^2}}. \blacktriangleright$$

▷ **Пример 5.8.** Найти математическое ожидание и дисперсию случайной величины $Y = 2 - 3\sin X$, если плотность вероятности случайной величины X есть $\varphi(x) = \dfrac{1}{2}\cos x$ на отрезке $[-\pi/2, \pi/2]$.

Р е ш е н и е. По формуле[1] **(5.57)**

$$a_y = M(Y) = \int_{-\pi/2}^{\pi/2} (2 - 3\sin x)\frac{1}{2}\cos x\,dx = 2.$$

Дисперсия $D(Y) = M(Y^2) - a_y^2$.

$$M(Y^2) = \int_{-\pi/2}^{\pi/2} (2 - 3\sin x)^2 \frac{1}{2}\cos x\,dx = 7 \text{ и } D(Y) = 7 - 2^2 = 3. \blacktriangleright$$

Из множества задач на составление закона распределения функции нескольких случайных величин важное для практики

[1] Вычисление интегралов предлагаем провести читателю самостоятельно.

значение имеет задача определения закона распределения с у м м ы двух случайных величин, т.е. закон распределения случайной величины $Z = X + Y$. В случае, если X и Y — независимые случайные величины, говорят о к о м п о з и ц и и (с в е р т к е) законов распределения.

В гл. 4, в частности, установлено, что сумма двух (и более) альтернативных случайных величин распределена по биномиальному закону; сумма двух случайных величин, распределенных по закону Пуассона, также распределена по закону Пуассона.

Рассмотрим композицию законов распределения двух непрерывных случайных величин. Пусть плотности вероятностей случайных величин X и Y равны соответственно $\varphi_1(x)$ и $\varphi_2(y)$.

Найдем сначала функцию распределения случайной величины Z:

$$F(z) = P(Z < z) = P(X + Y < z) = \iint_{D_z} \varphi(x, y)\,dx\,dy, \qquad \textbf{(5.58')}$$

где D_z — множество всех точек плоскости Oxy, координаты которых удовлетворяют неравенству $x+y < z$ (рис. 5.15), $\varphi(x, y)$ – совместная плотность двумерной случайной величины (X, Y). Так как X и Y — независимые случайные величины, то $\varphi(x, y) = \varphi_1(x)\varphi_2(y)$ и формула **(5.58')** примет вид:

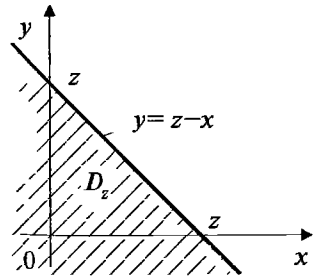


Рис. 5.15

$$F(z) = \int_{-\infty}^{+\infty}\int_{-\infty}^{z-x} \varphi_1(x)\varphi_2(y)\,dy =$$

$$= \int_{-\infty}^{+\infty} \varphi_1(x)\,dx \int_{-\infty}^{z-x} \varphi_2(y)\,dy.$$

Найдем плотность вероятности $\varphi(z)$:

$$\varphi(z) = F'(z) = \int_{-\infty}^{+\infty} \varphi_1(x)\,dx \left(\int_{-\infty}^{z-x} \varphi_2(y)\,dy\right)',$$

$$\varphi(z) = \int_{-\infty}^{+\infty} \varphi_1(x)\varphi_2(z-x)\,dx. \qquad \textbf{(5.59)}$$

Формулу **(5.59)** называют *формулой композиции* двух распределений или *формулой свертки*, которая в краткой записи имеет вид:

$$\varphi = \varphi_1 * \varphi_2.$$

▷ **Пример 5.9.** Найти закон распределения суммы двух случайных величин, распределенных равномерно на отрезке [0; 1].

Р е ш е н и е. Пусть $Z = X+Y$, где $\varphi_1(x) = 1$ при $0 \le x \le 1$ и $\varphi_2(y) = 1$ при $0 \le y \le 1$.

По формуле **(5.49)** плотность вероятности

$$\varphi(z) = \int_0^1 1 \cdot \varphi_2(z-x)\,dx = \int_0^1 \varphi_2(z-x)\,dx.$$

Если $z < 0$, то для $0 \le x \le 1$ $z - x < 0$; если $z > 2$, то для $0 \le x \le 1$ $z - x > 1$, следовательно, в этих случаях $\varphi_2(z-x) = 0$ и $\varphi(z) = 0$.

Пусть $0 \le z \le 2$. Подынтегральная функция $\varphi_2(z-x)$ будет отлична от нуля только для значений x, при которых $0 \le z - x \le 1$ или, что то же самое, при $z - 1 \le x \le z$.

Если $0 \le z \le 1$, то $\varphi(z) = \int_0^z 1 \cdot dx = z$.

Если $1 \le z \le 2$, то $\varphi(z) = \int_{z-1}^1 1 \cdot dx = 2 - z$.

Объединяя все случаи, получим:

$$\varphi(z) = \begin{cases} 0 & \text{при } z < 0,\ z > 2, \\ z & \text{при } 0 \le z \le 1, \\ 2-z & \text{при } 1 \le z \le 2. \end{cases} \qquad \textbf{(5.60)}$$

Закон распределения **(5.60)** называется *законом распределения Симпсона* или *законом равнобедренного треугольника* (рис. 5.16).

Вычисление $\varphi(z)$ можно было провести и иначе: вначале найти функцию распределения $F(z)$, а затем — ее производную, т.е. $\varphi(z) = F'(z)$. Преимущество такого подхода состоит в возможности использования геометрической интерпретации функции $F(z)$ как площади S_D области D — части квадрата (со стороной, равной 1), лежащей левее и ниже прямой $y = z - x$ (рис. 5.17).

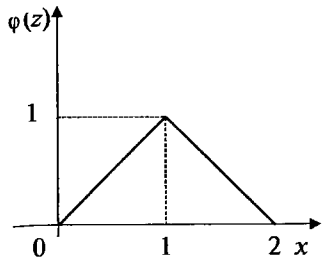


Рис. 5.16

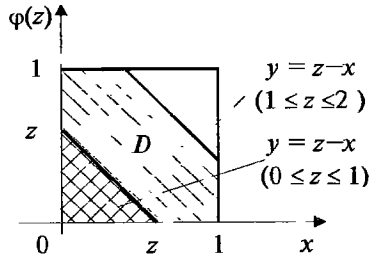


Рис. 5.17

Действительно (см. рис. 5.17), при $0 \le z \le 1$ $S_D = z^2/2$ (площадь заштрихованного треугольника со стороной z), а при $1 \le z \le 2$ $S_D = 1 - (2-z)^2/2$ (площадь квадрата без площади незаштрихованного треугольника, сторона которого, как нетрудно показать, равна $(2 - z)$. Следовательно,

$$F(z) = \begin{cases} 0 & \text{при } z < 0, \\ z^2/2 & \text{при } 0 \le z \le 1, \\ 1-(2-z)^2/2 & \text{при } 1 \le z \le 2, \\ 1 & \text{при } z > 2 \end{cases}$$

и выражение **(5.60)** для $\varphi(z)$ получается дифференцированием $F(z)$. ▶

Композиция нормальных законов распределения также *имеет нормальное распределение.* Так, если X и Y — независимые нормально распределенные случайные величины, т.е. $X \sim N(a_x, \sigma_x^2)$, $Y \sim N(a_y, \sigma_y^2)$, то случайная величина $Z = X+Y$ также нормально распределена: $Z \sim N(a_x+a_y,\ \sigma_x^2+\sigma_y^2)$. В случае, если нормально распределенные случайные величины X и Y зависимы (коэффициент корреляции $(\rho \ne 0)$, то случайная величина $Z = X + Y$ по-прежнему нормально распределена с параметрами $a_z = a_x + a_y$, $\sigma_z^2 = \sigma_x^2 + \sigma_y^2 + 2\rho\sigma_x\sigma_y$.

Упражнения

5.10. Закон распределения двумерной дискретной случайной величины (X, Y) задан в табл. 5.3.

Таблица 5.3

x_i \ y_j	0	1	2	3
−1	0,02	0,03	0,09	0,01
0	0,04	0,20	0,16	0,10
1	0,05	0,10	0,15	0,05

Найти: а) законы распределения одномерных случайных величин X и Y; б) условные законы распределения случайной величины X при условии $Y = 2$ и случайной величины Y при условии $X = 1$; в) вероятность $P(Y > X)$.

5.11. Рассматривается двумерная случайная величина (X, Y), где X — поставка сырья, Y — поступление требования на него. Известно, что поступление сырья и поступление требования на него могут произойти в любой день месяца (30 дней) с равной вероятностью. Определить: а) выражение совместной плотности и функции распределения двумерной случайной величины (X, Y); б) плотности вероятности и функции распределения одномерных составляющих X и Y; в) зависимы или независимы X и Y; г) вероятности того, что поставка сырья произойдет до и после поступления требования.

5.12. Двумерная случайная величина (X, Y) распределена равномерно внутри квадрата R с центром в начале координат. Стороны квадрата равны $\sqrt{2}$ и составляют углы 45° с осями координат. Определить: а) выражение совместной плотности двумерной случайной величины (X, Y); б) плотности вероятности одномерных составляющих X и Y; в) их условные плотности; г) зависимы или независимы X и Y.

5.13. Даны плотности вероятности независимых составляющих двумерной случайной величины (X, Y):

$$\varphi_1(x) = \begin{cases} 0 & \text{при } x < 0, \\ 5e^{-5x} & \text{при } x > 0; \end{cases} \qquad \varphi_2(y) = \begin{cases} 0 & \text{при } y < 0, \\ 2e^{-2y} & \text{при } y > 0. \end{cases}$$

Найти выражение совместной плотности и функции распределения двумерной случайной величины.

В примерах **5.14—5.16** определить: а) ковариацию и коэффициент корреляции случайных величин X и Y; б) коррелированы или некоррелированы эти случайные величины.

5.14. Использовать данные примера **5.10**.

5.15. Использовать данные примера **5.11**.

5.16. Использовать данные примера **5.12**.

5.17. Случайная величина X распределена на всей числовой оси с плотностью вероятности $\varphi(x) = 0{,}5e^{-|x|}$. Найти плотность вероятности случайной величины $Y = X^2$ и ее математическое ожидание.

5.18. Найти закон распределения суммы двух независимых случайных величин, каждая из которых распределена по стандартному нормальному закону, т.е. $N(0;1)$.

5.19. Двумерная случайная величина определяется следующим образом. Если при подбрасывании игральной кости выпадает четное число очков, то $X = 1$, в противном случае $X = 0$; $Y = 1$, когда число очков кратно трем, в противном случае $Y = 0$. Найти: а) законы распределения двумерной случайной величины (X, Y) и ее одномерных составляющих; б) условные законы распределения X и Y.

5.20. Двумерная случайная величина (X, Y) распределена с постоянной совместной плотностью внутри квадрата $OABC$, где $O(0;0)$, $A(0;1)$, $B(1;1)$, $C(1;0)$. Найти выражение совместной плотности и функции распределения двумерной случайной величины (X, Y).

5.21. Поверхность распределения двумерной случайной величины (X, Y) представляет прямой круговой конус, основанием которого служит круг с центром в начале координат и с радиусом 1. Вне этого круга совместная плотность двумерной случайной величины (X, Y) равна нулю. Найти выражения совместной плотности $\varphi(x, y)$, плотностей вероятностей одномерных составляющих $\varphi_1(x)$, $\varphi_2(y)$, условных плотностей $\varphi_x(y)$, $\varphi_y(x)$. Выяснить, являются ли случайные величины X и Y: зависимыми; коррелированными.

5.22. Двумерная случайная величина (X, Y) распределена по закону

$$\varphi(x,y)=\frac{A}{1+x^2+x^2y^2+y^2}.$$

Найти: а) коэффициент A; б) вероятность попадания случайной величины (X, Y) в пределы квадрата, центр которого совпадает с началом координат, а стороны параллельны осям координат и имеют длину 2. Установить, являются ли величины X и Y зависимыми; найти $\varphi_1(x)$, $\varphi_2(y)$.

5.23. Совместная плотность двумерной случайной величины (X, Y) имеет вид

$$\varphi(x,y)=\begin{cases}C(x^2+xy+y^2) & \text{при } 0\le x\le 1,\ 0\le y\le 1,\\ 0 & \text{в остальных случаях.}\end{cases}$$

Найти: а) постоянную C; б) плотности вероятности одномерных составляющих; в) их условные плотности; г) числовые характеристики a_x, a_y, $D(X)$, $D(Y)$, ρ.

5.24. Найти совместную плотность двумерной случайной величины (X, Y) и вероятность ее попадания в область D — прямоугольник, ограниченный прямыми $x=1$, $x=2$, $y=3$, $y=5$, если известна ее функция распределения (X, Y):

$$F(x,y)=\begin{cases}1-2^{-x}-2^{-y}+2^{-x-y} & \text{при } x\ge 0,\ y\ge 0,\\ 0 & \text{при } x<0 \text{ или } y<0.\end{cases}$$

5.25. Задана совместная плотность двумерной случайной величины (X, Y): $\varphi(x,y)=\dfrac{20}{\pi^2(16+x^2)(25+y^2)}$. Найти функцию распределения $F(x, y)$.

5.26. Имеются независимые случайные величины X, Y. Случайная величина X распределена по нормальному закону с параметрами $a_x=0$, $\sigma_x^2=1/2$. Случайная величина Y распределена равномерно на интервале (0;1). Найти выражения совместной плотности и функции распределения двумерной случайной величины (X, Y).

5.27. Совместная плотность двумерной случайной величины (X, Y) задана формулой

$$\varphi(x,y)=\frac{1}{1{,}6\pi}e^{-\frac{1}{1{,}28}[(x-2)^2-1{,}2(x-2)(y+3)+(y+3)^2]}$$

Найти a_x, a_y, σ_x^2, σ_y^2, ρ.

5.28. Независимые случайные величины X, Y распределены по нормальным законам с параметрами $a_x=2$, $a_y=-3$, $\sigma_x^2=1$, $\sigma_y^2=4$. Найти вероятности событий:

а) $(X<a_x)$ $(Y<a_y)$; б) $Y<X-5$; в) $(|X|<1)(|Y|<2)$.

5.29. Задана плотность вероятности $\varphi(x)$ случайной величины X, принимающей только положительные значения. Найти плотность вероятности случайной величины Y, если: а) $Y=e^{-X}$; б) $Y=\ln X$; в) $Y=X^3$; г) $Y=1/X^2$; д) $Y=\sqrt{X}$.

5.30. Случайная величина X равномерно распределена в интервале $(-\pi/2;\ \pi/2)$. Найти плотность вероятности случайной величины $Y=\sin X$.

5.31. Случайная величина распределена по закону Релея с плотностью вероятности $\varphi(x)=\begin{cases}xe^{-x^2/2} & \text{при } x>0,\\ 0 & \text{при } x<0.\end{cases}$ Найти закон распределения случайной величины $Y=e^{-X^2}$

5.32. Случайная величина X распределена по закону Коши с плотностью вероятности $\varphi(x)=\dfrac{1}{\pi(1+x^2)}$ $(-\infty<x<+\infty)$.

Найти плотность вероятности обратной величины $Y=1/X$.

5.33. Дискретная случайная величина X имеет ряд распределения

x_i	-1	0	1	2
p_i	0,2	0,1	0,3	0,4

Найти математическое ожидание и дисперсию случайной величины $Y=2^X$.

5.34. Имеются две случайные величины X и Y, связанные соотношением $Y=2-3X$. Числовые характеристики случайной величины X заданы $a_x=-1$; $D(X)=4$. Найти: а) математическое ожидание и дисперсию случайной величины Y; б) ковариацию и коэффициент корреляции случайной величин X и Y.

5.35. Случайная величина X задана плотностью вероятности $\varphi(x)=\cos x$ в интервале $(0,\ \pi/2)$; вне этого интервала $\varphi(x)=0$. Найти математическое ожидание случайной величины $Y=X^2$.

5.36. Случайная величина X распределена с постоянной плотностью вероятности в интервале (1;2) и нулевой плотностью вне этого интервала. Найти математическое ожидание и дисперсию случайной величины $Y = \frac{1}{X}$.

5.37. Непрерывная случайная величина X распределена в интервале (0;1) по закону с плотностью вероятности

$$\varphi(x) = \begin{cases} 2x & \text{при } x \in [0;1], \\ 0 & \text{при } x \notin [0;1]. \end{cases}$$

Найти математическое ожидание и дисперсию случайной величины $Y = X^2$.

5.38. Непрерывная случайная величина распределена по показательному закону с параметром $\lambda = 2$. Найти математическое ожидание и дисперсию случайной величины $Y = e^{-X}$.

5.39. Случайная величина X распределена по нормальному закону с параметрами $a = 0$, $\sigma^2 = 5$. Найти математическое ожидание случайной величины $Y = 1 - 3X^2 + 4X^3$.

5.40. Имеются две независимые случайные величины X и Y. Величина X распределена по нормальному закону с параметрами $a_x = 1$, $\sigma_x^2 = 4$. Величина Y распределена равномерно в интервале (0;2). Найти: а) $M(X - Y)$, $D(X - Y)$; б) $M(X^2)$, $M(Y^2)$.

5.41. Даны плотности вероятности независимых равномерно распределенных случайных величин X и Y: $\varphi_1(x) = 0{,}5$ в интервале (0; 2), вне этого интервала $\varphi_1(x) = 0$; $\varphi_2(y) = 0{,}5$ в интервале (0;2), вне этого интервала $\varphi_2(y) = 0$. Найти функцию распределения и плотность вероятности случайной величины $Z = X + Y$.

5.42. Независимые нормально распределенные случайные величины X и Y заданы плотностями вероятности $\varphi_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, $\varphi_2(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$. Найти композицию этих законов, т.е. плотность вероятности случайной величины $Z = X + Y$, и числовые характеристики a_z и σ_z^2.

Глава 6 Закон больших чисел и предельные теоремы

Под ***законом больших чисел*** в широком смысле понимается *общий принцип*, согласно которому, по формулировке академика А.Н. Колмогорова, *совокупное действие большого числа случайных факторов приводит* (при некоторых весьма общих условиях) к *результату, почти не зависящему от случая*. Другими словами, *при большом числе случайных величин их средний результат перестает быть случайным и может быть предсказан с большой степенью определенности.*

Под законом больших чисел в узком смысле понимается ряд математических теорем, в каждой из которых для тех или иных условий устанавливается факт приближения средних характеристик большого числа испытаний к некоторым определенным постоянным. Прежде чем перейти к этим теоремам, рассмотрим неравенства Маркова и Чебышева.

6.1. Неравенство Маркова (лемма Чебышева)

Теорема. *Если случайная величина X принимает только неотрицательные значения и имеет математическое ожидание, то для любого положительного числа A верно неравенство*

$$P(x > A) \le \frac{M(X)}{A}. \qquad (6.1)$$

□ Доказательство проведем для дискретной случайной величины X. Расположим ее значения в порядке возрастания, из которых часть значений $x_1, x_2, \ldots, x_k$ будут не более числа A, а другая часть — $x_{k+1}, \ldots, x_n$ будут больше A, т.е.

$$x_1 \le A, x_2 \le A, \ldots, x_k \le A; x_{k+1} > A, \ldots, x_n > A \text{ (рис. 6.1).}$$

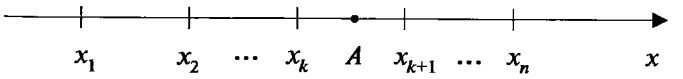


Рис. 6.1

Запишем выражение для математического ожидания $M(X)$:

$$x_1p_1+x_2p_2+...+x_kp_k+x_{k+1}p_{k+1}+...+x_np_n=M(X),$$

где p_1, p_2,..., p_n — вероятности того, что случайная величина X примет значения соответственно x_1, x_2,..., x_n.

Отбрасывая первые k неотрицательных слагаемых (напомним, что все $x_i \geq 0$), получим

$$x_{k+1}p_{k+1}+...+x_np_n \leq M(X). \qquad (6.2)$$

Заменяя в неравенстве (6.2) значения $x_{k+1},...,x_n$ меньшим числом A, получим более сильное неравенство

$$A(p_{k+1}+...+p_n) \leq M(X) \text{ или } p_{k+1}+...+p_n \leq \frac{M(X)}{A}.$$

Сумма вероятностей в левой части полученного неравенства представляет собой сумму вероятностей событий $X=x_{k+1},...,\ X=x_n$, т.е. вероятность события $X>A$. Поэтому $P(X>A) \leq \frac{M(X)}{A}$. ■

Так как события $X > A$ и $X \leq A$ противоположные, то заменяя $P(X>A)$ выражением $1-P(X \leq A)$, придем к другой форме неравенства Маркова:

$$P(X \leq A) \geq 1-\frac{M(X)}{A}. \qquad (6.3)$$

Неравенство Маркова применимо к любым неотрицательным случайным величинам.

▷ **Пример 6.1.** Среднее количество вызовов, поступающих на коммутатор завода в течение часа, равно 300. Оценить вероятность того, что в течение следующего часа число вызовов на коммутатор: а) превысит 400; б) будет не более 500.

Решение. а) По условию $M(X) = 300$. По формуле (6.1) $P(X>400) \leq \frac{300}{400}$, т.е. вероятность того, что число вызовов превысит 400, будет не более 0,75.

б) По формуле (6.3) $P(X \leq 500) \geq 1-\frac{300}{500}=0{,}4$, т.е. вероятность того, что число вызовов не более 500, будет не менее 0,4. ▶

▷ **Пример 6.2.** Сумма всех вкладов в отделение банка составляет 2 млн руб., а вероятность того, что случайно взятый вклад не превысит 10 тыс руб., равна 0,6. Что можно сказать о числе вкладчиков?

Решение. Пусть X — размер случайно взятого вклада, а n — число всех вкладов. Тогда из условия задачи следует, что средний размер вклада $M(X)=\frac{2000}{n}$ (тыс. руб.). Согласно неравенству Маркова (6.3): $P(X \leq 10) \geq 1-\frac{M(X)}{10}$ или $P(X \leq 10) \geq 1-\frac{2000}{10n}$.

Учитывая, что $P(X \leq 10)=0{,}6$, получим $1-\frac{200}{n} \leq 0{,}6$, откуда $n \leq 500$, т.е. число вкладчиков не более 500. ▶

6.2. Неравенство Чебышева

Теорема. *Для любой случайной величины, имеющей математическое ожидание и дисперсию, справедливо неравенство Чебышева:*

$$P(|X-a|>\varepsilon) \leq \frac{D(X)}{\varepsilon^2}, \qquad (6.4)$$

где $a = M(X)$, $\varepsilon > 0$.

□ Применим неравенство Маркова в форме (6.1) к случайной величине $X'=(X-a)^2$, взяв в качестве положительного числа $A=\varepsilon^2$. Получим

$$P[(X-a)^2 > \varepsilon^2] \leq \frac{M(X-a)^2}{\varepsilon^2}. \qquad (6.5)$$

Так как неравенство $(X-a)^2>\varepsilon^2$ равносильно неравенству $|X-a|>\varepsilon$, а $M(X-a)^2$ есть дисперсия случайной величины X, то из неравенства (6.5) получаем доказываемое неравенство (6.4). ■

Учитывая, что события $|X-a|>\varepsilon$ и $|X-a| \leq \varepsilon$ противоположны, неравенство Чебышева можно записать и в другой форме:

$$P(|X-a| \leq \varepsilon) \geq 1-\frac{D(X)}{\varepsilon^2}. \qquad (6.6)$$

Неравенство Чебышева применимо для любых случайных величин. В форме (6.4) оно устанавливает верхнюю границу, а в форме (6.6) — нижнюю границу вероятности рассматриваемого события.

Запишем неравенство Чебышева в форме **(6.6)** для некоторых случайных величин:

а) *для случайной величины $X = m$, имеющей биномиальный закон распределения с математическим ожиданием $a = M(X) = np$ и дисперсией $D(X) = npq$* (см. § 4.1):

$$P\left(|m-np|\le\varepsilon\right)\ge 1-\frac{npq}{\varepsilon^2}; \qquad \textbf{(6.7)}$$

б) *для частости $\frac{m}{n}$ события в n независимых испытаниях, в каждом из которых оно может произойти с одной и той же вероятностью $a = M\left(\frac{m}{n}\right) = p$, и имеющей дисперсию $D\left(\frac{m}{n}\right)=\frac{pq}{n}$:*

$$P\left(\left|\frac{m}{n}-p\right|\le\varepsilon\right)\ge 1-\frac{pq}{n\varepsilon^2}. \qquad \textbf{(6.8)}$$

▷ **Пример 6.3.** Средний расход воды на животноводческой ферме составляет 1000 *л* в день, а среднее квадратичное отклонение этой случайной величины не превышает 200 *л*. Оценить вероятность того, что расход воды на ферме в любой выбранный день не превзойдет 2000 *л*, используя: а) неравенство Маркова; б) неравенство Чебышева.

Р е ш е н и е. а) Пусть X — расход воды на животноводческой ферме (*л*). По условию $M(X) = 1000$. Используя неравенство Маркова **(6.3)**, получим $P(X\le 2000)\ge 1-\frac{1000}{2000}=0{,}5$, т.е. н е м е н е е, чем 0,5.

б) Дисперсия $D(X)=\sigma^2\le 200^2$. Так как границы интервала $0\le X\le 2000$ симметричны относительно математического ожидания $M(X) = 1000$, то для оценки вероятности искомого события можно применить неравенство Чебышева[1] **(6.6)**:

$$P(X\le 2000)=P(0\le X\le 2000)=$$
$$=P\left(|X-1000|\le 1000\right)\ge 1-\frac{200^2}{1000^2}=0{,}96,$$

[1] Берем в качестве дисперсии $D(X)$ ее максимальное значение, равное 200^2, что позволяет найти оценку вероятности искомого события для любых значений $D(X)\le 200^2$.

т.е. н е м е н е е, чем 0,96. В данной задаче оценку вероятности события, найденную с помощью неравенства Маркова $(P\ge 0{,}5)$, удалось уточнить с помощью неравенства Чебышева $(P\ge 0{,}96)$. ▶

▷ **Пример 6.4.** Вероятность выхода с автомата стандартной детали равна 0,96. Оценить с помощью неравенства Чебышева вероятность того, что число бракованных среди 2000 деталей находится в границах от 60 до 100 (включительно). Уточнить вероятность того же события с помощью интегральной теоремы Муавра—Лапласа. Объяснить различие полученных результатов.

Р е ш е н и е. По условию вероятность того, что деталь бракованная, равна $p = 1-0{,}96 = 0{,}04$. Число бракованных деталей $X = m$ имеет биномиальный закон распределения, а его границы 60 и 100 симметричны относительно математического ожидания $a = M(X) = np = 2000\cdot 0{,}04 = 80$.

Следовательно, оценку вероятности искомого события

$$P(60\le m\le 100)=P(-20\le m-80\le 20)=P\left(|m-80|\le 20\right)$$

можно найти по формуле **(6.6)**:

$$P\left(|m-80|\le 20\right)\ge 1-\frac{2000\cdot 0{,}04\cdot 0{,}96}{20^2}=1-\frac{76{,}8}{400}=0{,}808,$$

т.е. н е м е н е е чем 0,808.

Применяя следствие **(2.13)** интегральной теоремы Муавра—Лапласа, получим

$$P\left(|m-80|\le 20\right)\approx\Phi\left(\frac{20}{\sqrt{76{,}8}}\right)=\Phi(2{,}28)=0{,}979,$$

т.е. вероятность искомого события приближенно р а в н а 0,979.

Полученный результат $P\approx 0{,}979$ не противоречит оценке, найденной с помощью неравенства Чебышева — $P\ge 0{,}808$. Различие результатов объясняется тем, что неравенство Чебышева дает лишь нижнюю границу оценки вероятности искомого события для л ю б о й случайной величины, а интегральная теорема Муавра—Лапласа дает достаточно точное значение самой вероятности P (тем точнее, чем больше n), так как она применима лишь для случайной величины, имеющей о п р е д е л е н н ы й, а именно — биномиальный закон распределения. ▶

▷ **Пример 6.5.** Оценить вероятность того, что отклонение любой случайной величины от ее математического ожидания будет не более трех средних квадратических отклонений (по абсолютной величине) — (*правило трех сигм*).

Р е ш е н и е. По формуле (**6.6**), учитывая, что $D(X)=\sigma^2$, получим:

$$P\left(|X-a|\le 3\sigma\right)\ge 1-\frac{\sigma^2}{(3\sigma)^2}=\frac{8}{9}=0{,}889,$$

т.е. н е м е н е е, чем 0,889. Напомним, что для нормального закона правило трех сигм выполняется с вероятностью P, равной 0,9973, т.е. $P = 0{,}9973$. Можно показать, что для равномерного закона распределения $P = 1$, для показательного — $P = 0{,}9827$ и т.д. Таким образом, правило трех сигм (с достаточно большой вероятностью его выполнения) применимо для большинства случайных величин, встречающихся на практике. ▶

▷ **Пример 6.6.** По данным примера **2.8** с помощью неравенства Чебышева оценить вероятность того, что из 1000 новорожденных доля доживших до 50 лет будет отличаться от вероятности этого события не более, чем на 0,04 (по абсолютной величине).

Р е ш е н и е. Полагая $n = 1000$, $p = 0{,}87$, $q = 0{,}13$, по формуле (**6.7**)

$$P\left(\left|\frac{m}{n}-p\right|\le 0{,}04\right)\ge 1-\frac{0{,}87\cdot 0{,}13}{1000\cdot 0{,}04^2}=0{,}929,$$

т.е. н е м е н е е, чем 0,929. (Напомним, что в примере **2.8***б* было получено достаточно точное значение вероятности этого события при использовании следствия из интегральной теоремы Муавра—Лапласа, равное 0,9998; различие результатов объясняется так же, как и в примере **6.4**. ▶

З а м е ч а н и е. Если математическое ожидание $M(X) > A$ или дисперсия случайной величины $D(X) > \varepsilon^2$, то правые части неравенств Маркова и Чебышева в форме соответственно (**6.3**) и (**6.6**) будут отрицательными, а в форме (**6.1**) и (**6.4**) будут больше 1. Это означает, что применение указанных неравенств в этих случаях приведет к тривиальному результату: вероятность события больше отрицательного числа либо меньше числа, превосходящего 1. Но такой вывод очевиден и без использования данных

неравенств. Естественно, это обстоятельство снижает значение неравенств Маркова и Чебышева при решении практических задач, однако не умаляет их теоретического значения.

6.3. Теорема Чебышева

Теорема Чебышева. *Если дисперсии n независимых случайных величин X_1, X_2,..., X_n ограничены одной и той же постоянной, то при неограниченном увеличении числа n средняя арифметическая случайных величин сходится по вероятности к средней арифметической их математических ожиданий a_1, a_2,...,a_n,* т.е.

$$\lim_{n\to\infty} P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-\frac{a_1+a_2+\ldots+a_n}{n}\right|\le\varepsilon\right)=1 \qquad (6.9)$$

или

$$\frac{\sum_{i=1}^{n}X_i}{n}\xrightarrow[n\to\infty]{\mathcal{P}}\frac{\sum_{i=1}^{n}a_i}{n}. \qquad (6.10)$$

□ Вначале докажем формулу (**6.9**), затем выясним смысл формулировки «сходимость по вероятности». По условию

$$M(X_1)=a_1,\ M(X_2)=a_2,\ldots,\ M(X_n)=a_n,$$

$$D(X_1)\le C, D(X_2)\le C,\ldots, D(X_n)\le C,$$

где C — постоянное число.

Получим неравенство Чебышева в форме (**6.6**) для средней арифметической случайных величин, т.е. для

$$X=\frac{X_1+X_2+\ldots+X_n}{n}.$$

Найдем математическое ожидание $M(X)$ и оценку дисперсии $D(X)$:

$$M(X)=M\left(\frac{X_1+X_2+\ldots+X_n}{n}\right)=$$

$$=\frac{1}{n}\left[M(X_1)+M(X_2)+\ldots+M(X_n)\right]=\frac{a_1+a_2+\ldots+a_n}{n};$$

$$D(X) = D\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) =$$

$$= \frac{1}{n^2}\left[D(X_1) + D(X_2) + \ldots + D(X_n)\right] \le \frac{1}{n^2}\left(\underbrace{C + C + \ldots + C}_{n \text{ раз}}\right) = \frac{nC}{n^2} = \frac{C}{n}.$$

(Здесь использованы свойства математического ожидания и дисперсии и, в частности, то, что случайные величины X_1, X_2,...,X_n независимы, а следовательно, дисперсия их суммы равна сумме дисперсий.)

Запишем неравенство **(6.6)** для случайной величины $X=(X_1+X_2+\ldots+X_n)/n$:

$$P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{a_1 + a_2 + \ldots + a_n}{n}\right| \le \varepsilon\right) \ge 1 - \frac{D(X)}{\varepsilon^2}. \qquad \textbf{(6.11)}$$

Так как по доказанному $D(X) \le \frac{C}{n}$, то

$$1 - \frac{D(X)}{\varepsilon^2} \ge 1 - \frac{C/n}{\varepsilon^2} = 1 - \frac{C}{n\varepsilon^2},$$

и от неравенства **(6.11)** перейдем к более сильному неравенству:

$$P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{a_1 + a_2 + \ldots + a_n}{n}\right| \le \varepsilon\right) \ge 1 - \frac{C}{n\varepsilon^2}. \qquad \textbf{(6.12)}$$

В пределе при $n \to \infty$ величина $\frac{C}{n\varepsilon^2}$ стремится к нулю, и получим доказываемую формулу **(6.9)**. ■

Выясним теперь смысл формулировки *«сходимость по вероятности»* и записи ее содержания в виде **(6.10)**. Понятие предела переменной величины $X\left(\lim\limits_{n\to\infty} X = a \text{ или } X \to a \text{ при } n \to \infty\right)$ означает, что начиная с некоторого момента ее изменения для любого (даже сколь угодно малого) числа $\varepsilon > 0$ будет верно неравенство $|X - a| < \varepsilon$. В круглых скобках выражения **(6.9)** содержится аналогичное выражение[1]

[1] Записываем его кратко с помощью знаков суммирования.

$$\left|\left(\sum_{i=1}^{n} X_i\right)/n - \left(\sum_{i=1}^{n} a_i\right)/n\right| < \varepsilon,$$

где $\left(\sum\limits_{i=1}^{n} X_i\right)/n$ — случайная величина, а $\left(\sum\limits_{i=1}^{n} a_i\right)/n$ — постоянное число.

Однако из **(6.9)** вовсе не следует, что это неравенство будет выполняться всегда, начиная с некоторого момента изменения $\left(\sum\limits_{i=1}^{n} X_i\right)/n$. Так как $\left(\sum\limits_{i=1}^{n} X_i\right)/n$ — случайная величина, то возможно, что в отдельных случаях неравенство выполняться не будет. Однако с увеличением числа n вероятность неравенства $\left|\frac{\sum\limits_{i=1}^{n} X_i}{n} - \frac{\sum\limits_{i=1}^{n} a_i}{n}\right| \le \varepsilon$ стремится к 1, т.е. это неравенство будет выполняться в подавляющем числе случаев. Другими словами, при достаточно больших n выполнение рассматриваемого неравенства является событием *практически достоверным*, а неравенства противоположного смысла — *практически невозможным*.

Таким образом, стремление $\left(\sum\limits_{i=1}^{n} X_i\right)/n$ к $\left(\sum\limits_{i=1}^{n} a_i\right)/n$ следует понимать не как категорическое утверждение, а как утверждение, верность которого гарантируется с вероятностью, сколь угодно близкой к 1 при $n \to \infty$. Это обстоятельство и отражено в формулировке теоремы «сходится по вероятности» и в записи **(6.10)** обозначением $\xrightarrow[n\to\infty]{\mathcal{P}}$.

Подчеркнем смысл теоремы Чебышева. При большом числе n случайных величин X_1, X_2,...,X_n практически достоверно, что их средняя $X=\left(\sum\limits_{i=1}^{n} X_i\right)/n$ — величина случайная, как угодно мало отличается от неслучайной величины $\left(\sum\limits_{i=1}^{n} a_i\right)/n$, т.е. практически перестает быть случайной.

Следствие. *Если независимые случайные величины X_1, X_2,...,X_n имеют одинаковые математические ожидания, равные a, а их*

дисперсии ограничены одной и той же постоянной, то неравенство **(6.12)** *и формулы* **(6.9)**, **(6.10)** *примут вид*:

$$P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-a\right|\le\varepsilon\right)\ge 1-\frac{C}{n\varepsilon^2}, \qquad \textbf{(6.13)}$$

$$\lim_{n\to\infty} P\left(\left|\frac{X_1+X_2+\ldots+X_n}{n}-a\right|\le\varepsilon\right)=1, \qquad \textbf{(6.14)}$$

или

$$\frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow[n\to\infty]{P} a. \qquad \textbf{(6.15)}$$

□ Формулы **(6.13)**—**(6.15)** следуют из формул **(6.12)**, **(6.9)** и **(6.10)**, так как

$$M(X)=M\left(\frac{X_1+X_2+\ldots+X_n}{n}\right)=\frac{1}{n}\left[M(X_1)+M(X_2)+\ldots+M(X_n)\right]=$$

$$=\frac{1}{n}\left(\underbrace{a+a+\ldots+a}_{n\text{ раз}}\right)=\frac{na}{n}=a. \blacksquare$$

Теорема Чебышева и ее следствие имеют большое практическое значение. Например, страховой компании необходимо установить размер страхового взноса, который должен уплачивать страхователь; при этом страховая компания обязуется выплатить при наступлении страхового случая определенную страховую сумму. Рассматривая частоту/убытки страхователя при наступлении страхового случая как величину случайную и обладая известной статистикой таких случаев, можно определить среднее число/средние убытки при наступлении страховых случаев, которое на основании теоремы Чебышева с большой степенью уверенности можно считать величиной почти не случайной. Тогда на основании этих данных и предполагаемой страховой суммы определяется размер страхового взноса. Без учета действия закона больших чисел (теоремы Чебышева) возможны существенные убытки страховой компании (при занижении размера страхового взноса), либо потеря привлекательности страховых услуг (при завышении размера взноса).

Другой пример. Если надо измерить некоторую величину, истинное значение которой равно a, проводят n независимых измерений этой величины. Пусть результат каждого измерения — случайная величина X_i ($i = 1,2,\ldots, n$). Если при измерениях отсутствуют систематические погрешности (искажающие результат измерения в одну и ту же сторону), то естественно предположить, что $M(X_i) = a$ при любом i. Тогда на основании следствия из теоремы Чебышева средняя арифметическая результатов n измерений $\left(\sum_{i=1}^{n} X_i\right)/n$ сходится по вероятности к истинному значению a. Этим обосновывается выбор средней арифметической в качестве меры истинного значения a.

Если все измерения проводятся с одинаковой точностью характеризуемой дисперсией $D(X_i)=\sigma^2$, то дисперсия их средней

$$D\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)=\frac{1}{n^2}D\left(\sum_{i=1}^{n} X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n} D(X_i)=\frac{1}{n^2}\left(n\sigma^2\right)=\frac{\sigma^2}{n},$$

а ее среднее квадратическое отклонение равно $\sigma/\sqrt{n}$. Полученное отношение, известное под названием «*правила корня из n*», говорит о том, что средний ожидаемый разброс средней n измерений в $\sqrt{n}$ раз меньше разброса каждого измерения. Таким образом, увеличивая число измерений, можно как угодно уменьшать влияние случайных погрешностей (но не систематических), т.е. увеличивать точность определения истинного значения a.

Замечание. Если измерительный прибор имеет точность δ (например, δ — половина ширины деления равномерной шкалы прибора, по которой производится отсчет), то указанным выше способом нельзя рассчитывать получить точность измерения величины a бóльшую, чем δ. Каждое измерение дает результат с неопределенностью δ и, очевидно, их средняя арифметическая будет обладать той же неопределенностью δ. Таким образом, стремиться посредством закона больших чисел получить значение a с большей степенью точности, чем позволяет прибор при отдельном измерении, является заблуждением.

▷ **Пример 6.7.** Для определения средней продолжительности горения электроламп в партии из 200 одинаковых ящиков было взято на выборку по одной лампе из каждого ящика. Оценить вероятность того, что средняя продолжительность горения отобранных 200 электроламп отличается от средней продолжительности

горения ламп во всей партии не более чем на 5 ч (по абсолютной величине), если известно, что среднее квадратическое отклонение продолжительности горения ламп в каждом ящике меньше 7 ч.

Р е ш е н и е. Пусть X_i — продолжительность горения электролампы, взятой из i-го ящика (ч). По условию дисперсия $D(X_i) < 7^2 = 49$. Очевидно, что средняя продолжительность горения отобранных ламп равна $(X_1+X_2+...+X_{200})/200$, а средняя продолжительность горения ламп во всей партии $(M(X_1) + M(X_2) + ...+M(X_{200}))/200=(a_1+a_2+...+a_{200})/200$.

Тогда вероятность искомого события по формуле **(6.12)**:

$$P\left(\left|\frac{X_1+X_2+...+X_{200}}{200}-\frac{a_1+a_2+...+a_{200}}{200}\right|\le 5\right)\ge 1-\frac{49}{200\cdot 5^2}\approx 0{,}9902,$$

т.е. н е м е н е е, чем 0,9902. ▶

▷ **Пример 6.8.** Сколько надо провести измерений данной величины, чтобы с вероятностью не менее 0,95 гарантировать отклонение средней арифметической этих измерений от истинного значения величины не более, чем на 1 (по абсолютной величине), если среднее квадратическое отклонение каждого из измерений не превосходит 5?

Р е ш е н и е. Пусть X_i — результат i-го измерения ($i = 1,2,...,n$); a — истинное значение величины, т.е. $M(X_i) = a$ при любом i.

Необходимо найти n, при котором

$$P\left(\left|\frac{X_1+X_2+...+X_n}{n}-a\right|\le 1\right)\ge 0{,}95.$$

В соответствии с **(6.12)** данное неравенство будет выполняться, если

$$1-\frac{C}{n\varepsilon^2}=1-\frac{5^2}{n\cdot 1^2}\ge 0{,}95, \text{ откуда } \frac{25}{n}\le 0{,}05$$

и $n\ge\frac{25}{0{,}05}=500$, т.е. потребуется н е м е н е е 500 измерений. ▶

6.4. Теорема Бернулли

Теорема Бернулли. *Частость события в n повторных независимых испытаниях, в каждом из которых оно может произойти с одной и той же вероятностью p, при неограниченном увеличении*

числа n сходится по вероятности к вероятности p этого события в отдельном испытании:

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n}-p\right|\le\varepsilon\right)=1 \qquad \textbf{(6.16)}$$

или

$$\frac{m}{n}\xrightarrow[n\to\infty]{\mathcal{P}} p. \qquad \textbf{(6.17)}$$

□ Заключение теоремы **(6.16)** непосредственно вытекает из неравенства Чебышева для частости события **(6.8)** при $n\to\infty$. ■

С м ы с л теоремы Бернулли состоит в том, что при большом числе n повторных независимых испытаний практически достоверно, что частость (или статистическая вероятность) события m/n — величина с л у ч а й н а я, как угодно мало отличается от н е с л у ч а й н о й величины p — вероятности события, т.е. практически перестает быть случайной.

З а м е ч а н и е. Теорема Бернулли является следствием теоремы Чебышева, ибо частость события можно представить как среднюю арифметическую n независимых альтернативных случайных величин, имеющих один и тот же закон распределения **(4.4)** (см. § 4.1). Доказательство теоремы (более громоздкое) возможно и без ссылки на теорему (неравенство) Чебышева. Исторически эта теорема была доказана намного раньше более общей теоремы Чебышева.

Теорема Бернулли дает теоретическое обоснование замены неизвестной вероятности события его частостью, или статистической вероятностью (см. § 1.3), полученной в n повторных независимых испытаниях, проводимых при одном и том же комплексе условий. Так, например, если вероятность рождения мальчика нам не известна, то в качестве ее значения мы можем принять частость (статистическую вероятность) этого события, которая, как известно по многолетним статистическим данным, составляет приближенно 0,515.

Теорема Бернулли является звеном, позволяющим связать формальное аксиоматическое определение вероятности (см. § 1.12) с эмпирическим (опытным) законом постоянства относительной частоты (см. § 1.3). Теорема дает возможность обосновать широкое применение на практике вероятностных методов исследования.

Непосредственным обобщением теоремы Бернулли является теорема Пуассона, когда вероятности события в каждом испытании различны.

Теорема Пуассона. *Частость события в n повторных независимых испытаниях, в каждом из которых оно может произойти*

соответственно с вероятностями $p_1, p_2,..., p_n$, *при неограниченном увеличении числа n сходится по вероятности к средней арифметической вероятностей события в отдельных испытаниях*, т.е.

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n} - \frac{p_1 + p_2 + \ldots + p_n}{n}\right| \le \varepsilon\right) = 1 \qquad \textbf{(6.18)}$$

или

$$\frac{m}{n} \underset{n\to\infty}{\xrightarrow{\mathcal{P}}} \frac{\sum_{i=1}^{n} p_i}{n}. \qquad \textbf{(6.19)}$$

□ Теорема Пуассона непосредственно вытекает из теоремы Чебышева, если в качестве случайных величин $X_1, X_2,..., X_n$ рассматривать альтернативные случайные величины, имеющие законы распределения вида **(4.4)** с параметрами $p_1, p_2,..., p_n$. Так как математические ожидания случайных величин $X_1, X_2,..., X_n$ равны соответственно $p_1, p_2,..., p_n$, а их дисперсии $p_1q_1, p_2q_2,...,$ p_nq_n (см. § 4.1) ограничены одним числом[1], то формула **(6.18)** непосредственно вытекает из формулы **(6.9)**. ■

Важная роль закона больших чисел в теоретическом обосновании методов математической статистики и ее приложений обусловила проведение ряда исследований, направленных на изучение общих условий применимости этого закона к последовательности случайных величин. Так, в теореме Маркова доказана справедливость предельного равенства **(6.15)** для зависимых случайных величин X_i $(i = 1,2,...,n)$ при условии

$$\lim_{n\to\infty} \frac{1}{n^2} D\left(\sum_{i=1}^{n} X_i\right) = 0.$$

Например, температура воздуха в некоторой местности X_i $(i = 1,2,...,365)$ каждый день года — величины случайные, подверженные существенным колебаниям в течение года, причем зависимые, ибо на погоду каждого дня, очевидно, заметно влияет погода предыдущих дней. Однако среднегодовая температура $\left(\sum_{i=1}^{365} X_i\right)/365$ почти не меняется для данной местности в течение многих лет, являясь практически неслучайной, предопределенной.

[1] Легко показать, что для любого i имеем

$$p_iq_i = p_i(1 - p_i) = -p_i^2 + p_i = -(p_i - 0{,}5)^2 + 0{,}25 \le 0{,}25.$$

Нахождение общих условий, выполнение которых обязательно влечет за собой статистическую устойчивость средних, представляет непреходящую научную ценность исследований в области закона больших чисел.

Помимо различных форм закона больших чисел в теории вероятностей имеются еще разные формы так называемого «*усиленного закона больших чисел*», где показывается не «сходимость по вероятности», а «сходимость с вероятностью 1» различных средних случайных величин к неслучайным средним. Однако этот усиленный закон представляет больше интерес в теоретических исследованиях и не столь важен для его приложений в экономике.

6.5. Центральная предельная теорема

Рассмотренный выше закон больших чисел устанавливает факт приближения средней большого числа случайных величин к определенным постоянным. Но этим не ограничиваются закономерности, возникающие в результате суммарного действия случайных величин. Оказывается, что при некоторых условиях совокупное действие случайных величин приводит к определенному, а именно — к нормальному закону распределения.

Центральная предельная теорема представляет собой группу теорем, посвященных установлению условий, при которых возникает нормальный закон распределения. Среди этих теорем важнейшее место приналлежит теореме Ляпунова.

Теорема Ляпунова. *Если* $X_1, X_2,..., X_n$ — *независимые случайные величины, у каждой из которых существует математическое ожидание* $M(X_i)=a$, *дисперсия* $D(X_i) = \sigma^2$, *абсолютный центральный момент третьего порядка* $M\left(|X_i - a_i|^3\right) = m_i$ и

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} m_i}{\left(\sum_{i=1}^{n} \sigma_i^{\,2}\right)^{3/2}} = 0, \qquad \textbf{(6.20)}$$

то закон распределения суммы $Y_n = X_1 + X_2 + \ldots + X_n$ *при* $n \to \infty$ *неограниченно приближается к нормальному с математическим ожиданием* $\sum_{i=1}^{n} a_i$ *и дисперсией* $\sum_{i=1}^{n} \sigma_i^2$.

Теорему принимаем без доказательства.

Неограниченное приближение закона распределения суммы $Y_n = \sum_{i=1}^{n} X_i$ к нормальному закону при $n \to \infty$ в соответствии со свойствами нормального закона означает, что

$$\lim_{n\to\infty} P\left(\left|\frac{Y_n - \sum_{i=1}^{n} a_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}\right| \le z\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-t^2/2}dt = \frac{1}{2} + \frac{1}{2}\Phi(z), \qquad \textbf{(6.21)}$$

где $\Phi(z)$ — функция Лапласа (**2.11**).

Смысл условия (**6.20**) состоит в том, чтобы в сумме $Y_n = \sum_{i=1}^{n} X_i$ не было слагаемых, влияние которых на рассеяние Y_n подавляюще велико по сравнению с влиянием всех остальных, а также не должно быть большого числа случайных слагаемых, влияние которых очень мало по сравнению с суммарным влиянием остальных. Таким образом, *удельный вес каждого отдельного слагаемого должен стремиться к нулю при увеличении числа слагаемых.*

Так, например, потребление электроэнергии для бытовых нужд за месяц в каждой квартире многоквартирного дома можно представить в виде n различных случайных величин. Если потребление электроэнергии в каждой квартире по своему значению резко не выделяется среди остальных, то на основании теоремы Ляпунова можно считать, что потребление электроэнергии всего дома, т.е. сумма n независимых случайных величин будет случайной величиной, имеющей приближенно нормальный закон распределения. Если, например, в одном из помещений дома разместится вычислительный центр, у которого уровень потребления электроэнергии несравнимо выше, чем в каждой квартире для бытовых нужд, то вывод о приближенно нормальном распределении потребления электроэнергии всего дома будет неправомерен, так как нарушено условие (**6.20**), ибо потребление электроэнергии вычислительного центра будет играть превалирующую роль в образовании всей суммы потребления.

Другой пример. При устойчивом и отлаженном режиме работы станков, однородного обрабатываемого материала и т.д.

варьирование качества продукции принимает форму нормального закона распределения в силу того, что производственная погрешность представляет собой результат суммарного действия большого числа случайных величин: погрешности станка, инструмента, рабочего и т.д.

Следствие. *Если $X_1, X_2,...,X_n$ — независимые случайные величины, у которых существуют равные математические ожидания $M(X_i)=a$, дисперсии $D(X_i)=\sigma^2$ и абсолютные центральные моменты третьего порядка $M\left(\left|X_i - a_i\right|^3\right) = m_i$ $(i = 1, 2,...,n)$, то закон распределения суммы $Y_n=X_1+X_2+...+X_n$ при $n \to \infty$ неограниченно приближается к нормальному закону.*

□ Доказательство сводится к проверке условия (**6.20**):

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} m}{\left(\sum_{i=1}^{n} \sigma_i^2\right)^{3/2}} = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} m}{\left(\sum_{i=1}^{n} \sigma^2\right)^{3/2}} = \lim_{n\to\infty} \frac{mn}{\left(n\sigma^2\right)^{3/2}} = \lim_{n\to\infty} \frac{m}{\sigma^3\sqrt{n}} = 0;$$

следовательно, имеет место и равенство (**6.21**). ■

В частности, *если все случайные величины X_i одинаково распределены, то закон распределения их суммы неограниченно приближается к нормальному закону при $n \to \infty$.*

Проиллюстрируем это утверждение на примере суммирования независимых случайных величин, имеющих равномерное распределение на интервале (0, 1). Кривая распределения одной такой случайной величины показана на рис. 6.2*а*. На рис. 6.2*б* показана плотность вероятности суммы двух таких случайных величин (см. пример **5.9**), а на рис. 6.2*в* — плотность вероятности суммы трех таких случайных величин (ее график состоит из трех отрезков парабол на интервалах (0, 1), (1, 2) и (2, 3) и по виду уже напоминает нормальную кривую).

Если сложить шесть таких случайных величин, то получится случайная величина с плотностью вероятности, практически не отличающейся от нормальной.

Теперь у нас имеется возможность д о к а з а т ь ***локальную и интегральную теоремы Муавра—Лапласа*** (см. § 2.3).

Рассмотрим случайную величину $Z=\dfrac{m-np}{\sqrt{npq}}$, где $X=m$ — число появлений события в n независимых испытаниях, в каждом из которых оно может появиться с одной и той же вероятностью p, т.е. $X = m$ — случайная величина, имеющая биномиальный закон распределения, для которого математическое ожидание $M(X)=np$ и дисперсия $D(X)=npq$.

Случайная величина Z, так же как случайная величина X, вообще говоря, дискретна, но при большом числе n испытаний ее значения расположены на оси абсцисс так тесно, что ее можно рассматривать как непрерывную с плотностью вероятности $\varphi(z)$.

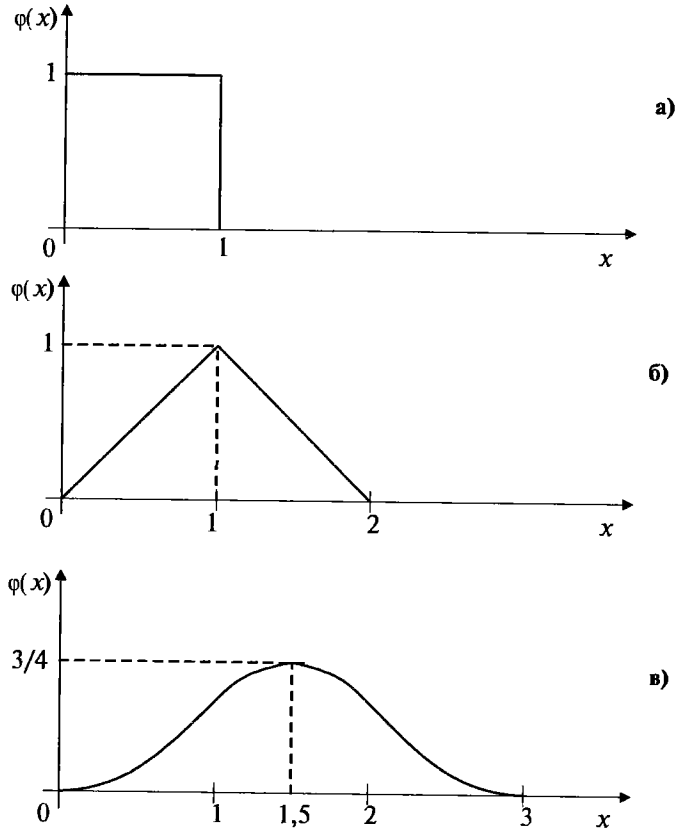


Рис. 6.2

Найдем числовые характеристики случайной величины Z, используя свойства математического ожидания и дисперсии:

$$a = M(Z) = (M(X) - np)/\sqrt{npq} = (np - np)/\sqrt{npq} = 0,$$

$$D(Z) = (D(X) - 0) / \left(\sqrt{npq}\right)^2 = npq / npq = 1.$$

В силу того, что случайная величина X представляет собой сумму независимых альтернативных случайных величин (см. § 4.1), случайная величина Z представляет также сумму независимых, одинаково распределенных случайных величин и, следовательно, на основании центральной предельной теоремы при большом числе n имеет распределение, близкое к нормальному закону с параметрами $a = 0$, $\sigma^2 = 1$. Используя свойство **(4.32)** нормального закона, с учетом **(4.33)** получим

$$P(z_1 \le Z \le z_2) \approx \frac{1}{2}\left[\Phi(z_2) - \Phi(z_1)\right]. \qquad \textbf{(6.22)}$$

Полагая $z_1 = \dfrac{a-np}{\sqrt{npq}}$, $z_2 = \dfrac{b-np}{\sqrt{npq}}$, с учетом того, что $Z = \dfrac{m-np}{\sqrt{npq}}$, получаем, что двойное неравенство в скобках равносильно неравенству $a \le m \le b$. В результате из формулы **(6.22)** получим *интегральную формулу Муавра—Лапласа* **(2.10)**:

$$P(a \le m \le b) \approx \frac{1}{2}\left[\Phi(z_2) - \Phi(z_1)\right]. \qquad \textbf{(6.23)}$$

Вероятность $P_{m,n}$ того, что событие A произойдет m раз в n независимых испытаниях, можно приближенно записать в виде:

$$P_{m,n} \approx P_n(m \le X \le m + \Delta m).$$

Чем меньше Δm, тем точнее приближенное равенство. Минимальное (целое) $\Delta m = 1$. Поэтому, учитывая **(6.23)** и **(6.22)**, можно записать:

$$P_{m,n} \approx \frac{1}{2}\left[\Phi(z_2) - \Phi(z_1)\right] = P(z_1 \le Z \le z_2), \qquad \textbf{(6.24)}$$

где $z_1 = \dfrac{m-np}{\sqrt{npq}}$, $z_2 = \dfrac{(m+1)-np}{\sqrt{npq}}$.

При малых Δz имеем

$$P(z + \Delta z) \approx \varphi(z)\Delta z, \qquad \textbf{(6.25)}$$

где $\varphi(z)$ — плотность стандартной нормально распределенной случайной величины с параметрами $a = 0$, $\sigma^2 = 1$, т.е.

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad (6.26)$$

Полагая $z_1 = z$, $\Delta z = z_2 - z_1 = \frac{(m+1)-np}{\sqrt{npq}} - \frac{m-np}{\sqrt{npq}} = \frac{1}{\sqrt{npq}}$, из формулы **(6.25)** с учетом **(6.24)** получим *локальную формулу Муавра—Лапласа* **(2.7)**:

$$P_{m,n} \approx \frac{1}{\sqrt{npq}} \varphi(z). \qquad (6.27)$$

Замечание. Необходимо соблюдать известную осторожность, применяя центральную предельную теорему в статистических исследованиях. Так, если сумма $\sum_{i=1}^{n} X_i$ при $n \to \infty$ всегда имеет нормальный закон распределения, то скорость сходимости к нему существенно зависит от типа распределения ее слагаемых. Так, например, как отмечено выше, при суммировании равномерно распределенных случайных величин уже при 6—10 слагаемых можно добиться достаточной близости к нормальному закону, в то время как для достижения той же близости при суммировании χ^2-распределенных случайных слагаемых понадобится более 100 слагаемых.

Опираясь на центральную предельную теорему, можно утверждать, что рассмотренные в гл. 4 случайные величины, имеющие законы распределения — биномиальный, Пуассона, гипергеометрический, χ^2 («хи-квадрат»), t (Стьюдента), при $n \to \infty$ распределены асимптотически нормально.

Упражнения

6.9. Среднее изменение курса акции компании в течение одних биржевых торгов составляет 0,3%. Оценить вероятность того, что на ближайших торгах курс изменится более, чем на 3%.

6.10. Отделение банка обслуживает в среднем 100 клиентов в день. Оценить вероятность того, что сегодня в отделении банка будет обслужено: а) не более 200 клиентов; б) более 150 клиентов.

6.11. Электростанция обслуживает сеть на 1600 электроламп, вероятность включения каждой из которых вечером равна 0,9. Оценить с помощью неравенства Чебышева вероятность того, что число ламп, включенных в сеть вечером, отличается от своего математического ожидания не более чем на 100 (по абсолютной величине). Найти вероятность того же события, используя следствие из интегральной теоремы Муавра—Лапласа.

6.12. Вероятность того, что акции, переданные на депозит, будут востребованы, равна 0,08. Оценить с помощью неравенства Чебышева вероятность того, что среди 1000 клиентов от 70 до 90 востребуют свои акции.

6.13. Среднее значение длины детали 50 см, а дисперсия — 0,1. Используя неравенство Чебышева, оценить вероятность того, что случайно взятая деталь окажется по длине не менее 49,5 и не более 50,5 см. Уточнить вероятность того же события, если известно, что длина случайно взятой детали имеет нормальный закон распределения.

6.14. Оценить вероятность того, что отклонение любой случайной величины от ее математического ожидания будет не более двух средних квадратических отклонений (по абсолютной величине).

6.15. В течение времени t эксплуатируются 500 приборов. Каждый прибор имеет надежность 0,98 и выходит из строя независимо от других. Оценить с помощью неравенства Чебышева вероятность того, что доля надежных приборов отличается от 0,98 не более чем на 0,1 (по абсолютной величине).

6.16. Вероятность сдачи в срок всех экзаменов студентом факультета равна 0,7. С помощью неравенства Чебышева оценить вероятность того, что доля сдавших в срок все экзамены из 2000 студентов заключена в границах от 0,66 до 0,74.

6.17. Бензоколонка N заправляет легковые и грузовые автомобили. Вероятность того, что проезжающий легковой автомобиль подъедет на заправку, равна 0,3. С помощью неравенства Чебышева найти границы, в которых с вероятностью, не меньшей 0,79, находится доля заправившихся в течение 2 ч легковых автомобилей, если за это время всего заправилось 100 автомобилей.

6.18. В среднем 10% работоспособного населения некоторого региона — безработные. Оценить с помощью неравенства Чебышева вероятность того, что уровень безработицы среди обследованных 10 000 работоспособных жителей города будет в пределах от 9 до 11% (включительно).

6.19. Выход цыплят в инкубаторе составляет в среднем 70% числа заложенных яиц. Сколько нужно заложить яиц, чтобы с вероятностью, не меньшей 0,95, ожидать, что отклонение числа вылупившихся цыплят от математического ожидания их не превышало 50 (по абсолютной величине)? Решить задачу с помощью: а) неравенства Чебышева; б) интегральной теоремы Муавра—Лапласа.

6.20. Опыт работы страховой компании показывает, что страховой случай приходится примерно на каждый пятый договор. Оценить с помощью неравенства Чебышева необходимое количество договоров, которые следует заключить, чтобы с вероятностью 0,9 можно было утверждать, что доля страховых случаев отклонится от 0,1 не более чем на 0,01 (по абсолютной величине). Уточнить ответ с помощью следствия из интегральной теоремы Муавра—Лапласа.

6.21. В целях контроля из партии в 100 ящиков взяли по одной детали из каждого ящика и измерили их длину. Требуется оценить вероятность того, что вычисленная по данным выборки средняя длина детали отличается от средней длины детали во всей партии не более чем на 0,3 мм, если известно, что среднее квадратическое отклонение не превышает 0,8 мм.

6.22. Сколько нужно произвести измерений, чтобы с вероятностью, равной 0,9973, утверждать, что погрешность средней арифметической результатов этих измерений не превысит 0,01, если измерение характеризуется средним квадратическим отклонением, равным 0,03?

Глава 7 Элементы теории случайных процессов и теории массового обслуживания

Теория случайных процессов (*случайных функций*) — это раздел математической науки, изучающий закономерности случайных явлений в д и н а м и к е их развития.

7.1. Определение случайного процесса и его характеристики

Понятие случайного процесса является обобщением понятия случайной величины, рассмотренной в гл. 3.

О п р е д е л е н и е. ***Случайным процессом*** $X(t)$ *называется процесс, значение которого при любом значении аргумента* t *является случайной величиной.*

Другими словами, случайный процесс представляет собой функцию, которая в результате испытания может принять тот или иной конкретный вид, неизвестный заранее. При фиксированном $t = t_0$ $X(t_0)$ представляет собой обычную случайную величину, т.е. с е ч е н и е случайного процесса в момент t_0.

Аналогично тому, как в гл. 3 записана случайная величина в виде функции элементарного события ω, появляющегося в результате испытания, случайный процесс можно записать в виде функции двух переменных $X(t, \omega)$, где $\omega \in \Omega$, $t \in T$, $X(t, \omega) \in \Xi$ и ω — элементарное событие, Ω — пространство элементарных событий, T — множество значений аргумента t, Ξ — множество возможных значений случайного процесса $X(t, \omega)$.

Реализацией случайного процесса $X(t, \omega)$ называется неслучайная функция $x(t)$, в которую превращается случайный процесс $X(t)$ в результате испытания (при фиксированном ω), т.е. конкретный вид, принимаемый случайным процессом $X(t)$, его *траектория*.

Таким образом, *случайный процесс* $X(t, \omega)$ *совмещает в себе черты случайной величины и функции.* Если зафиксировать значение аргумента t, случайный процесс превращается в обычную случайную величину, если зафиксировать ω, то в результате каждого испытания он превращается в обычную неслучайную функцию. В дальнейшем изложении опустим аргумент ω, но он будет подразумеваться по умолчанию.

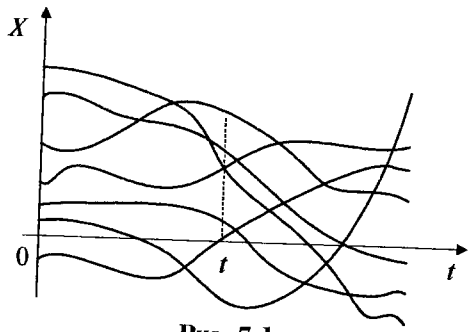


Рис. 7.1

На рис. 7.1 изображено несколько реализаций некоторого случайного процесса. Пусть сечение этого процесса при данном t является непрерывной случайной величиной. Тогда случайный процесс $X(t)$ при данном t определяется плотностью вероятности $\varphi(x, t)$.

Очевидно, что плотность $\varphi(x, t)$ не является исчерпывающим описанием случайного процесса $X(t)$, ибо она не выражает зависимости между его сечениями в разные моменты времени.

Случайный процесс $X(t)$ представляет собой совокупность всех сечений при всевозможных значениях t, поэтому для его описания необходимо рассматривать многомерную случайную величину $(X(t_1), X(t_2),..., X(t_n))$, состоящую из всех сечений этого процесса. В принципе таких сечений бесконечно много, но для описания случайного процесса удается часто обойтись относительно небольшим количеством сечений.

Говорят, что случайный процесс имеет *порядок* n, если он полностью определяется плотностью совместного распределения $\varphi(x_1, x_2,..., x_n; t_1, t_2,..., t_n)$ n произвольных сечений процесса, т.е. плотностью n-мерной случайной величины $(X(t_1), X(t_2),..., X(t_n))$, где $X(t_i)$ — сечение случайного процесса $X(t)$ в момент времени t_i, $i = 1,2,...,n$.

Как и случайная величина, случайный процесс может быть описан числовыми характеристиками.

Математическим ожиданием *случайного процесса $X(t)$ называется неслучайная функция $a_x(t)$, которая при любом значении переменной t равна математическому ожиданию соответствующего сечения случайного процесса $X(t)$, т.е. $a_x(t) = M[X(t)]$.*

Дисперсией *случайного процесса $X(t)$ называется неслучайная функция $D_x(t)$, при любом значении переменной t равная дисперсии соответствующего сечения случайного процесса $X(t)$,* т.е. $D_x(t) = D[X(t)]$.

Средним квадратическим отклонением $\sigma_x(t)$ *случайного процесса $X(t)$ называется арифметическое значение корня квадратного из его дисперсии, т.е.* $\sigma_x(t) = \sqrt{D_x(t)}$.

Математическое ожидание случайного процесса характеризует среднюю траекторию всех возможных его реализаций, а

его дисперсия или среднее квадратическое отклонение — разброс реализаций относительно средней траектории.

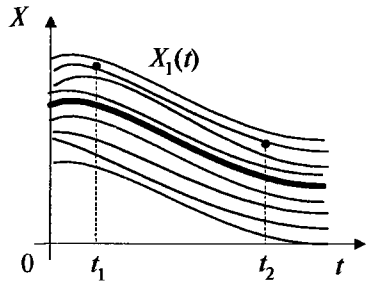


Рис. 7.2

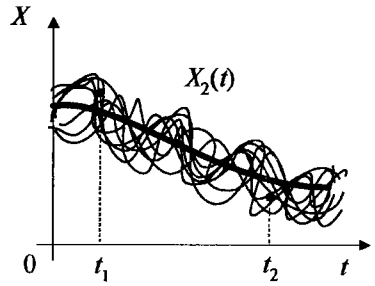


Рис. 7.3

Введенных выше характеристик случайного процесса оказывается недостаточно, так как они определяются только одномерным законом распределения. На рис. 7.2 и 7.3 изображены два случайных процесса $X_1(t)$ и $X_2(t)$ с примерно одинаковыми математическими ожиданиями и дисперсиями.

Если для случайного процесса $X_1(t)$ (рис. 7.2) характерно медленное изменение значений реализаций с изменением t, то для случайного процесса $X_2(t)$ (рис. 7.3) это изменение происходит значительно быстрее. Другими словами, для случайного процесса $X_1(t)$ характерна тесная вероятностная зависимость между двумя его сечениями $X_1(t_1)$ и $X_1(t_2)$, в то время как для случайного процесса $X_2(t)$ эта зависимость между сечениями $X_2(t_1)$ и $X_2(t_2)$ практически отсутствует. Указанная зависимость между сечениями характеризуется корреляционной функцией.

Определение. ***Корреляционной функцией*** *случайного процесса $X(t)$ называется неслучайная функция*

$$K_x(t_1, t_2) = M[(X(t_1)-a_x(t_1))(X(t_2)-a_x(t_2))] \qquad (7.1)$$

двух переменных t_1 и t_2, которая при каждой паре переменных t_1 и t_2 равна ковариации соответствующих сечений $X(t_1)$ и $X(t_2)$ случайного процесса.

Очевидно, для случайного процесса $X_1(t)$ корреляционная функция $K_{x_1}(t_1,t_2)$ убывает по мере увеличения разности $t_2 - t_1$ значительно медленнее, чем $K_{x_2}(t_1,t_2)$ для случайного процесса $X_2(t)$.

Корреляционная функция $K_x(t_1,t_2)$ характеризует не только степень тесноты линейной зависимости между двумя сечения-

ми, но и разброс этих сечений относительно математического ожидания $a_x(t)$. Поэтому рассматривается также нормированная корреляционная функция случайного процесса.

Нормированной корреляционной функцией *случайного процесса* $X(t)$ *называется функция*

$$\rho_x(t_1,t_2)=\frac{K_x(t_1,t_2)}{\sigma_x(t_1)\sigma_x(t_2)}. \qquad (7.2)$$

▷ **Пример 7.1.** Случайный процесс определяется формулой $X(t) = X\cos\omega t$, где X — случайная величина. Найти основные характеристики этого процесса, если $M(X) = a$, $D(X) = \sigma^2$.

Р е ш е н и е. На основании свойств математического ожидания и дисперсии имеем:

$$a_x(t)=M(X\cos\omega t)=\cos\omega t\cdot M(X)=a\cos\omega t,$$

$$D_x(t)=D(X\cos\omega t)=\cos^2\omega t\cdot D(X)=\sigma^2\cos^2\omega t.$$

Корреляционную функцию найдем по формуле (**7.1**):

$$K_x(t_1,t_2)=M\left[(X\cos\omega t_1-a\cos\omega t_1)(X\cos\omega t_2-a\cos\omega t_2)\right]=$$

$$=\cos\omega t_1\cos\omega t_2\cdot M\left[(X-a)(X-a)\right]=\cos\omega t_1\cos\omega t_2\cdot D(X)=$$

$$=\sigma^2\cos\omega t_1\cos\omega t_2.$$

Нормированную корреляционную функцию найдем по формуле (**7.2**):

$$\rho_x(t_1,t_2)=\frac{\sigma^2\cos\omega t_1\cos\omega t_2}{(\sigma\cos\omega t_1)(\sigma\cos\omega t_2)}\equiv 1. \blacktriangleright$$

Случайные процессы можно классифицировать в зависимости от того, плавно или скачкообразно меняются состояния системы, в которой они протекают, конечно (счетно) или бесконечно множество этих состояний и т.п. Среди случайных процессов особое место принадлежит *марковскому случайному процессу* (§ 7.3). Но прежде познакомимся с основными понятиями теории массового обслуживания.

7.2. Основные понятия теории массового обслуживания

На практике часто приходится сталкиваться с системами, предназначенными для многоразового использования при решении однотипных задач. Возникающие при этом процессы получили название *процессов обслуживания*, а системы — *систем массового обслуживания* (СМО). Примерами таких систем являются телефонные системы, ремонтные мастерские, вычислительные комплексы, билетные кассы, магазины, парикмахерские и т.п.

Каждая СМО состоит из определенного числа обслуживающих единиц (приборов, устройств, пунктов, станций), которые будем называть *каналами* обслуживания. Каналами могут быть линии связи, рабочие точки, вычислительные машины, продавцы и др. По числу каналов СМО подразделяют на *одноканальные* и *многоканальные*.

Заявки поступают в СМО обычно не регулярно, а случайно, образуя так называемый *случайный поток заявок* (*требований*). Обслуживание заявок, вообще говоря, также продолжается какое-то случайное время. Случайный характер потока заявок и времени обслуживания приводит к тому, что СМО оказывается загруженной неравномерно: в какие-то периоды времени скапливается очень большое количество заявок (они либо становятся в очередь, либо покидают СМО необслуженными), в другие же периоды СМО работает с недогрузкой или простаивает.

Предметом теории массового обслуживания является построение математических моделей, связывающих заданные условия работы СМО (число каналов, их производительность, характер потока заявок и т.п.) с показателями эффективности СМО, описывающими ее способность справляться с потоками заявок.

В качестве *показателей эффективности* СМО используются: среднее[1] число заявок, обслуживаемых в единицу времени; среднее число заявок в очереди; среднее время ожидания обслуживания; вероятность отказа в обслуживании без ожидания; вероятность того, что число заявок в очереди превысит определенное значение, и т.п.

СМО делят на два основных типа (класса): СМО *с отказами* и СМО *с ожиданием* (*очередью*). В СМО с отказами заявка, поступившая в момент, когда все каналы заняты, получает отказ, покидает СМО и в дальнейшем процессе обслуживания не участвует (например, заявка на телефонный разговор в момент, когда все каналы заняты, получает отказ и покидает СМО необслуженной). В СМО с ожиданием заявка, пришедшая в момент, когда все каналы заняты, не уходит, а становится в очередь на обслуживание.

СМО с ожиданием подразделяются на разные виды в зависимости от того, как организована очередь: с ограниченной или

[1] Здесь и в дальнейшем средние величины понимаются как математические ожидания соответствующих случайных величин.

неограниченной длиной очереди, с ограниченным временем ожидания и т.п.

7.3. Понятие марковского случайного процесса

Процесс работы СМО представляет собой *случайный процесс*.

Процесс называется *процессом с дискретными состояниями*, если его возможные состояния S_1, S_2, S_3... можно заранее перечислить, а переход системы из состояния в состояние происходит мгновенно (скачком). Процесс называется *процессом с непрерывным временем*, если моменты возможных переходов системы из состояния в состояние не фиксированы заранее, а случайны.

Процесс работы СМО представляет собой случайный процесс с дискретными состояниями и непрерывным временем. Это означает, что состояние СМО меняется скачком в случайные моменты появления каких-то событий (например, прихода новой заявки, окончания обслуживания и т.п.).

Математический анализ работы СМО существенно упрощается, если процесс этой работы — марковский. Случайный процесс называется *марковским* или *случайным процессом без последействия*, если для любого момента времени t_0 вероятностные характеристики процесса в будущем зависят только от его состояния в данный момент t_0 и не зависят от того, когда и как система пришла в это состояние.

Пример марковского процесса: система S — счетчик в такси. Состояние системы в момент t характеризуется числом километров (десятых долей километров), пройденных автомобилем до данного момента. Пусть в момент t_0 счетчик показывает S_0. Вероятность того, что в момент $t > t_0$ счетчик покажет то или иное число километров (точнее, соответствующее число рублей) S_1, зависит от S_0, но не зависит от того, в какие моменты времени изменялись показания счетчика до момента t_0.

Многие процессы можно приближенно считать марковскими. Например, процесс игры в шахматы; система S — группа шахматных фигур. Состояние системы характеризуется числом фигур противника, сохранившихся на доске в момент t_0. Вероятность того, что в момент $t > t_0$ материальный перевес будет на стороне одного из противников, зависит в первую очередь от того, в каком состоянии находится система в данный момент t_0, а не от того, когда и в какой последовательности исчезли фигуры с доски до момента t_0.

В ряде случаев предысторией рассматриваемых процессов можно просто пренебречь и применять для их изучения марковские модели.

При анализе случайных процессов с дискретными состояниями удобно пользоваться геометрической схемой — так называемым *графом состояний*. Обычно состояния системы изображаются прямоугольниками (кружками), а возможные переходы из состояния в состояние — стрелками (ориентированными дугами), соединяющими состояния.

▷ **Пример 7.2.** Построить граф состояний следующего случайного процесса: устройство S состоит из двух узлов, каждый из которых в случайный момент времени может выйти из строя, после чего мгновенно начинается ремонт узла, продолжающийся заранее неизвестное случайное время.

Р е ш е н и е. Возможные состояния системы: S_0 — оба узла исправны; S_1 — первый узел ремонтируется, второй исправен; S_2 — второй узел ремонтируется, первый исправен; S_3 — оба узла ремонтируются. Граф системы приведен на рис. 7.4.

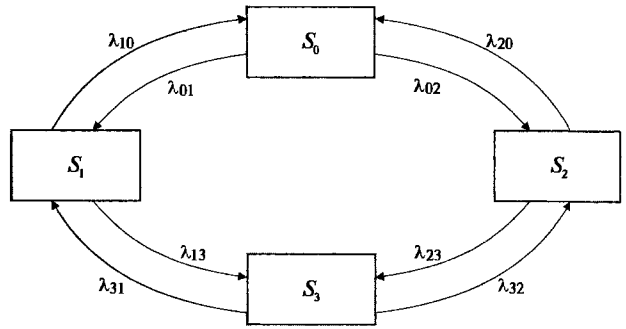


Рис. 7.4

Стрелка, направленная, например, из S_0 в S_1, означает переход системы в момент отказа первого узла, из S_1 в S_0 — переход в момент окончания ремонта этого узла.

На графе отсутствуют стрелки из S_0 в S_3 и из S_1 в S_2. Это объясняется тем, что выходы узлов из строя предполагаются независимыми друг от друга и, например, вероятностью одновременного выхода из строя двух узлов (переход из S_0 в S_3) или одновременного окончания ремонтов двух узлов (переход из S_3 в S_0) можно пренебречь. ▶

Для математического описания марковского случайного процесса с дискретными состояниями и непрерывным временем, протекающего в СМО, познакомимся с одним из важных понятий теории вероятностей — понятием потока событий.

7.4. Потоки событий

Под *потоком событий* понимается последовательность однородных событий, следующих одно за другим в какие-то случайные моменты времени (например, поток вызовов на телефонной станции, поток отказов ЭВМ, поток покупателей и т.п.).

Поток характеризуется *интенсивностью* λ — частотой появления событий или средним числом событий, поступающих в СМО в единицу времени.

Поток событий называется *регулярным*, если события следуют одно за другим через определенные равные промежутки времени. Например, поток изделий на конвейере сборочного цеха (с постоянной скоростью движения) является регулярным.

Поток событий называется *стационарным*, если его вероятностные характеристики не зависят от времени. В частности, интенсивность стационарного потока есть величина постоянная: $\lambda(t)=\lambda$. Например, поток автомобилей на городском проспекте не является стационарным в течение суток, но этот поток можно считать стационарным в определенное время суток, скажем, в часы пик. В этом случае фактическое число проходящих автомобилей в единицу времени (например, в каждую минуту) может заметно различаться, но среднее их число постоянно и не будет зависеть от времени.

Поток событий называется *потоком без последействия*, если для любых двух непересекающихся участков времени τ_1 и τ_2 число событий, попадающих на один из них, не зависит от числа событий, попадающих на другие. Например, поток пассажиров, входящих в метро, практически не имеет последействия. А, скажем, поток покупателей, отходящих с покупками от прилавка, уже имеет последействие (хотя бы потому, что интервал времени между отдельными покупателями не может быть меньше, чем минимальное время обслуживания каждого из них).

Поток событий называется *ординарным*, если вероятность попадания на малый (элементарный) участок времени Δt двух и более событий пренебрежимо мала по сравнению с вероятностью попадания одного события. Другими словами, поток событий ординарен, если события появляются в нем поодиночке, а

не группами. Например, поток поездов, подходящих к станции, ординарен, а поток вагонов не ординарен.

Поток событий называется ***простейшим*** *(**или стационарным пуассоновским**), если он одновременно стационарен, ординарен и не имеет последействия.* Название «простейший» объясняется тем, что СМО с простейшими потоками имеет наиболее простое математическое описание. Регулярный поток не является простейшим, так как обладает последействием: моменты появления событий в таком потоке жестко зафиксированы.

Простейший поток в качестве предельного возникает в теории случайных процессов столь же естественно, как в теории вероятностей нормальное распределение получается в качестве предельного для суммы случайных величин: *при наложении (суперпозиции) достаточно большого числа n независимых, стационарных и ординарных потоков (сравнимых между собой по интенсивностям λ_i ($i=1,2,\ldots, n$)) получается поток, близкий к простейшему с интенсивностью λ, равной сумме интенсивностей входящих потоков*, т.е.

$$\lambda=\sum_{i=1}^{n}\lambda_i\,. \qquad (7.3)$$

Рассмотрим на оси времени Ot (рис. 7.5) простейший поток событий как неограниченную последовательность случайных точек.

Рис. 7.5

Пусть случайная величина X выражает число событий (точек), попадающих на произвольный промежуток времени τ. Покажем, что *случайная величина X распределена по закону Пуассона.*

□ Разобьем мысленно временной промежуток τ на n равных элементарных отрезков $\Delta t = \tau/n$. Математическое ожидание числа событий, попадающих на элементарный отрезок Δt, очевидно, равно $\lambda\Delta t$, где λ — интенсивность потока. Согласно свойству ординарности потока можно пренебречь вероятностью попадания на элементарный отрезок двух и более событий.

Будем считать элементарный отрезок Δt «занятым», если в нем появилось событие потока, и «свободным», если не появи-

лось. Вероятность того, что отрезок $\Delta t = \tau/n$ окажется «занятым», равна $\lambda\Delta t \approx \lambda\tau/n$; вероятность того, что он окажется «пустым», равна $1 - \lambda\tau/n$ (чем меньше Δt, тем точнее равенства).

Число занятых элементарных отрезков, т.е. число X событий на всем временном промежутке τ, можно рассматривать как случайную величину, имеющую биномиальный закон распределения, т.е.

$$P(X = m) = C_n^m \left(\frac{\lambda\tau}{n}\right)^m \left(1 - \frac{\lambda\tau}{n}\right)^{n-m} \qquad (7.4)$$

с параметрами n и $p = \lambda\tau/n$.

(Необходимое для возникновения биномиального закона условие независимости испытаний, в данном случае — независимость n элементарных отрезков относительно события «отрезок занят», обеспечивается свойством отсутствия последействия потока).

При неограниченном увеличении числа элементарных отрезков Δt, т.е. при $n \to \infty$, $p = \dfrac{\lambda\tau}{n} \to 0$ и постоянном значении произведения $np = n\dfrac{\lambda\tau}{n} = \lambda\tau$, как отмечено в § 4.2, биномиальное распределение стремится к распределению Пуассона с параметром $\lambda\tau$:

$$P(X = m) = \frac{(\lambda\tau)^m}{m!} e^{-\lambda\tau}, \qquad (7.5)$$

для которого математическое ожидание случайной величины равно ее дисперсии: $a = \sigma^2 = \lambda\tau$.

В частности, вероятность того, что за время τ не произойдет ни одного события ($m = 0$), равна

$$P(X = 0) = e^{-\lambda\tau}. \blacksquare \qquad (7.6)$$

▷ **Пример 7.3.** На автоматическую телефонную станцию поступает простейший поток вызовов с интенсивностью λ=1,2 вызовов в минуту. Найти вероятность того, что за две минуты: а) не придет ни одного вызова; б) придет ровно один вызов; в) придет хотя бы один вызов.

Р е ш е н и е. а) Случайная величина X — число вызовов за две минуты — распределена по закону Пуассона с параметром $\lambda\tau$=1,2·2=2,4. Вероятность того, что вызовов не будет (m=0), по формуле (7.5):

$$P(X = 0) \approx e^{-2,4} = 0,091.$$

б) Вероятность одного вызова (m=1):

$$P(X = 1) \approx 2,4 \cdot 0,091 = 0,218.$$

в) Вероятность хотя бы одного вызова:

$$P(X \geq 1) = 1 - P(X = 0) = 1 - 0,091 = 0,909. \blacktriangleright$$

Найдем распределение интервала времени T между двумя произвольными соседними событиями простейшего потока.

В соответствии с (7.6) вероятность того, что на участке времени длиной t не появится ни одного из последующих событий, равна

$$P(T \geq t) = e^{-\lambda t}, \qquad (7.7)$$

а вероятность противоположного события, т.е. функция распределения случайной величины T, есть

$$F(t) = P(T < t) = 1 - e^{-\lambda t} \qquad (7.8)$$

Плотность вероятности случайной величины есть производная ее функции распределения (рис. 7.6), т.е.

$$\varphi(t) = F'(t) = \lambda e^{-\lambda t}. \qquad (7.9)$$

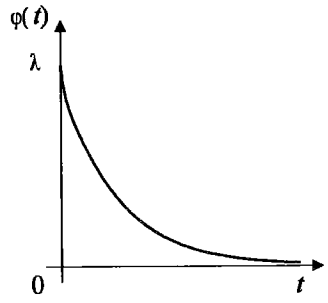


Рис. 7.6

Распределение, задаваемое плотностью вероятности (7.9) или функцией распределения (7.8), является *показательным* (*экспоненциальным*) (см. § 4.6). Таким образом, *интервал времени между двумя соседними произвольными событиями простейшего потока имеет показательное распределение*, для которого математическое ожидание равно среднему квадратическому отклонению случайной величины:

$$a = \sigma = \frac{1}{\lambda} \qquad (7.10)$$

и обратно по величине интенсивности потока λ.

Важнейшее свойство показательного распределения (присущее только показательному распределению) состоит в следующем: *если промежуток времени, распределенный по показательному закону, уже длился некоторое время* τ, *то это никак не влияет*

на закон распределения оставшейся части промежутка $(T-\tau)$: *он будет таким же, как и закон распределения всего промежутка* T (см. гл. 4, пример **4.7**).

Другими словами, для интервала времени T между двумя последовательными соседними событиями потока, имеющего показательное распределение, любые сведения о том, сколько времени протекал этот интервал, не влияют на закон распределения оставшейся части. Это свойство показательного закона представляет собой, в сущности, другую формулировку для «отсутствия последействия» — основного свойства простейшего потока.

Для простейшего потока с интенсивностью λ вероятность попадания на *элементарный* (*малый*) отрезок времени Δt хотя бы одного события потока равна, согласно (**7.8**),

$$P_{\Delta t} = P(T < \Delta t) = 1 - e^{-\lambda \Delta t} \approx \lambda \Delta t . \qquad (7.11)$$

(Эта приближенная формула, получаемая заменой функции $e^{-\lambda \Delta t}$ лишь двумя первыми членами ее разложения в ряд по степеням Δt, тем точнее, чем меньше Δt).

7.5. Уравнения Колмогорова. Предельные вероятности состояний

Рассмотрим математическое описание марковского случайного процесса с дискретными состояниями и непрерывным временем по данным примера **7.2**. Соответствующий граф состояний процесса изображен на рис. 7.4. Будем полагать, что все переходы системы из состояния S_i в S_j происходят под воздействием простейших потоков событий с интенсивностями λ_{ij} ($i, j = 0, 1, 2, 3$); так, переход системы из состояния S_0 в S_1 будет происходить под воздействием потока отказов первого узла, а обратный переход из состояния S_1 в S_0 — под воздействием потока «окончаний ремонтов» первого узла и т.п.

Граф состояний системы с проставленными у стрелок интенсивностями будем называть *размеченным* (см. рис. 7.4). Рассматриваемая система S имеет четыре возможных состояния: S_0, S_1, S_2, S_3.

Вероятностью i-го состояния называется вероятность $p_i(t)$ того, что в момент t система будет находиться в состоянии S_i. Очевидно, что для любого момента t сумма вероятностей всех состояний равна единице:

$$\sum_{i=0}^{3} p_i(t) = 1. \qquad (7.12)$$

Рассмотрим систему в момент t и, задав малый промежуток Δt, найдем вероятность $p_0(t + \Delta t)$ того, что система в момент $t+\Delta t$ будет находиться в состоянии S_0. Это достигается разными способами.

1. Система в момент t с вероятностью $p_0(t)$ находилась в состоянии S_0, а за время Δt не вышла из него.

Вывести систему из этого состояния (см. граф на рис. 7.4) можно суммарным простейшим потоком с интенсивностью $(\lambda_{01}+\lambda_{02})$, т.е. в соответствии с (**7.11**), с вероятностью, приближенно равной $(\lambda_{01}+\lambda_{02})\Delta t$. А вероятность того, что система не выйдет из состояния S_0, равна $[1-(\lambda_{01}+\lambda_{02})\Delta t]$. Вероятность того, что система будет находиться в состоянии S_0 и не выйдет из него за время Δt, равна по теореме умножения вероятностей:

$$p_0(t)\left[1 - (\lambda_{01} + \lambda_{02})\Delta t\right].$$

2. Система в момент t с вероятностями $p_1(t)$ (или $p_2(t)$) находилась в состоянии S_1 или S_2 и за время Δt перешла в состояние S_0.

Потоком интенсивностью λ_{10} (или λ_{20}) (см. рис. 7.4) система перейдет в состояние S_0 с вероятностью, приближенно равной $\lambda_{10}\Delta t$ (или $\lambda_{20}\Delta t$). Вероятность того, что система будет находиться в состоянии S_0, по этому способу равна $p_1(t)\lambda_{10}\Delta t$ (или $p_2(t)\lambda_{20}\Delta t$).

Применяя теорему сложения вероятностей, получим:

$$p_0(t+\Delta t) = p_1(t)\lambda_{10}\Delta t + p_2(t)\lambda_{20}\Delta t + p_0(t)\left[1 - (\lambda_{01} + \lambda_{02})\Delta t\right],$$

откуда

$$\frac{p_0(t+\Delta t) - p_0(t)}{\Delta t} = p_1(t)\lambda_{10} + p_2(t)\lambda_{20} - (\lambda_{01} + \lambda_{02})\, p_0(t).$$

Переходя к пределу при $\Delta t \to 0$ (приближенные равенства, связанные с применением формулы (**7.11**), перейдут в точные), получим в левой части уравнения производную $p'_0(t)$ (обозначим ее для простоты p'_0):

$$p'_0 = \lambda_{10}\, p_1 + \lambda_{20}\, p_2 - (\lambda_{01} + \lambda_{02})\, p_0 .$$

Получено дифференциальное уравнение первого порядка, т.е. уравнение, содержащее как саму неизвестную функцию, так и ее производную первого порядка.

Рассуждая аналогично для других состояний системы S, можно получить *систему дифференциальных уравнений Колмогорова для вероятностей состояний*:

$$\begin{cases} p_0' = \lambda_{10}p_1 + \lambda_{20}p_2 - (\lambda_{01} + \lambda_{02})p_0, \\ p_1' = \lambda_{01}p_0 + \lambda_{31}p_3 - (\lambda_{10} + \lambda_{13})p_1, \\ p_2' = \lambda_{02}p_0 + \lambda_{32}p_3 - (\lambda_{20} + \lambda_{23})p_2, \\ p_3' = \lambda_{13}p_1 + \lambda_{23}p_2 - (\lambda_{31} + \lambda_{32})p_3. \end{cases} \quad (7.13)$$

Сформулируем правило составления уравнений Колмогорова. *В левой части каждого из них стоит производная вероятности i-го состояния. В правой части — сумма произведений вероятностей всех состояний (из которых идут стрелки в данное состояние) на интенсивности соответствующих потоков событий минус суммарная интенсивность всех потоков, выводящих систему из данного состояния, умноженная на вероятность данного (i-го состояния)* — см. рис. 7.4.

В системе (7.13) независимых уравнений на единицу меньше общего числа уравнений. Поэтому для решения системы необходимо добавить уравнение (7.12).

Особенность решения дифференциальных уравнений вообще состоит в том, что требуется задавать так называемые начальные условия, в данном случае — вероятности состояний системы в начальный момент $t = 0$. Так, например, систему уравнений (7.13) естественно решать при условии, что в начальный момент оба узла исправны и система находилась в состоянии S_0, т.е. при начальных условиях $p_0(0)=1$, $p_1(0)=p_2(0)=p_3(0)=0$.

Уравнения Колмогорова дают возможность найти все вероятности состояний как *функции времени*. Особый интерес представляют вероятности системы $p_i(t)$ в *предельном стационарном режиме*, т.е. при $t \to \infty$, которые называются *предельными* (*финальными*) вероятностями состояний.

В теории случайных процессов доказывается, что *если число состояний системы конечно и из каждого из них можно (за конечное число шагов) перейти в любое другое состояние, то предельные вероятности существуют.*

Предельная вероятность состояния S_i имеет четкий смысл: она показывает *среднее относительное время пребывания системы в этом состоянии*. Например, если предельная вероятность состояния S_0, т.е. p_0=0,5, то это означает, что в среднем половину времени система находится в состоянии S_0.

Так как предельные вероятности постоянны, то, заменяя в уравнениях Колмогорова их производные нулевыми значениями, получим систему линейных алгебраических уравнений, описывающих стационарный режим. Для системы S с графом состояний, изображенном на рис. 7.4, такая система уравнений имеет вид:

$$\begin{cases} (\lambda_{01} + \lambda_{02})p_0 = \lambda_{10}p_1 + \lambda_{20}p_2, \\ (\lambda_{10} + \lambda_{13})p_1 = \lambda_{01}p_0 + \lambda_{31}p_3, \\ (\lambda_{20} + \lambda_{23})p_2 = \lambda_{02}p_0 + \lambda_{32}p_3, \\ (\lambda_{31} + \lambda_{32})p_3 = \lambda_{13}p_1 + \lambda_{23}p_2. \end{cases} \quad (7.14)$$

Систему (7.14) можно составить непосредственно по размеченному графу состояний, если руководствоваться правилом, согласно которому *слева в уравнениях стоит предельная вероятность данного состояния p_i, умноженная на суммарную интенсивность всех потоков, ведущих из данного состояния, а справа — сумма произведений интенсивностей всех потоков, входящих в i-е состояние, на вероятности тех состояний, из которых эти потоки исходят.*

▷ **Пример 7.4.** Найти предельные вероятности для системы S из примера 7.2, граф состояний которой приведен на рис. 7.4, при λ_{01}=1, λ_{02}=2, λ_{10}=2, λ_{13}=2, λ_{20}=3, λ_{23}=1, λ_{31}=3, λ_{32}=2.

Решение. Система алгебраических уравнений, описывающих стационарный режим для данной системы, имеет вид (7.14) или

$$\begin{cases} 3p_0 = 2p_1 + 3p_2, \\ 4p_1 = p_0 + 3p_3, \\ 4p_2 = 2p_0 + 2p_3, \\ p_0 + p_1 + p_2 + p_3 = 1. \end{cases} \quad (7.15)$$

(Здесь вместо одного «лишнего» уравнения системы (7.14) записали нормировочное условие (7.12).)

Решив систему (7.15), получим p_0=0,40, p_1=0,20, p_2=0,27, p_3=0,13, т.е. в предельном стационарном режиме система S в среднем 40% времени будет находиться в состоянии S_0 (оба узла исправны), 20% — в состоянии S_1 (первый узел ремонтируется, второй работает), 27% — в состоянии S_2 (второй узел ремонти-

руется, первый работает) и 13% времени — в состоянии S_3 (оба узла ремонтируются).▶

▷ **Пример 7.5.** Найти средний чистый доход от эксплуатации в стационарном режиме системы S в условиях примеров **7.2** и **7.4**, если известно, что в единицу времени исправная работа первого и второго узлов приносит доход соответственно в 10 и 6 ден. ед., а их ремонт требует затрат соответственно в 4 и 2 ден. ед. Оценить экономическую эффективность имеющейся возможности уменьшения вдвое среднего времени ремонта каждого из двух узлов, если при этом придется вдвое увеличить затраты на ремонт каждого узла (в единицу времени).

Р е ш е н и е. Из примера **7.4** следует, что в среднем первый узел исправно работает долю времени, равную $p_0+p_2=0{,}40+0{,}27=0{,}67$, а второй узел — $p_0+p_1=0{,}40+0{,}20=0{,}60$. В то же время первый узел находится в ремонте в среднем долю времени, равную $p_1+p_3=0{,}20+0{,}13=0{,}33$, а второй узел — $p_2+p_3=0{,}27+0{,}13=0{,}40$. Поэтому средний чистый доход в единицу времени от эксплуатации системы, т.е. разность между доходами и затратами, равен

$$D = 0{,}67\cdot 10+0{,}60\cdot 6-0{,}33\cdot 4-0{,}40\ 2 = 8{,}18 \text{ ден. ед.}$$

Уменьшение вдвое среднего времени ремонта каждого из узлов в соответствии с **(7.10)** будет означать увеличение вдвое интенсивностей потока «окончаний ремонтов» каждого узла, т.е. теперь $\lambda_{10}=4$, $\lambda_{20}=6$, $\lambda_{31}=6$, $\lambda_{32}=4$ и система линейных алгебраических уравнений **(7.14)**, описывающая стационарный режим системы S, вместе с нормировочным условием **(7.12)** примет вид[1]:

$$\begin{cases} 3p_0 = 4p_1 + 6p_2, \\ 6p_1 = p_0 + 6p_3, \\ 7p_2 = 2p_0 + 4p_3, \\ p_0 + p_1 + p_2 + p_3 = 1. \end{cases}$$

Решив систему, получим $p_0=0{,}60$, $p_1=0{,}15$, $p_2=0{,}20$, $p_3=0{,}05$.

Учитывая, что $p_0+p_2=0{,}60+0{,}20=0{,}80$, $p_0+p_1=0{,}60+0{,}15=0{,}75$, $p_1+p_3=0\ 15+0{,}05=0{,}20$ $p_2+p_3=0{,}20+0{,}05=0{,}25$, а затраты на ремонт первого и второго узлов составляют теперь соответственно 8 и 4 ден. ед., вычислим средний чистый доход в единицу времени:

$$D_1 = 0{,}80\cdot 10+0{,}75\cdot 6-0{,}20\cdot 8-0{,}25\cdot 4 = 9{,}9 \text{ ден. ед.}$$

Так как D_1 больше D (примерно на 20%), то экономическая целесообразность ускорения ремонтов узлов очевидна. ▶

[1] При записи системы **(7.14)** одно «лишнее» уравнение исключили.

7.6. Процессы гибели и размножения

В теории массового обслуживания широко распространен специальный класс случайных процессов — так называемые *процессы гибели и размножения*. Название это связано с рядом биологических задач, где этот процесс служит математической моделью изменения численности биологических популяций.

Граф состояний процесса гибели и размножения имеет вид, показанный на рис. 7.7.

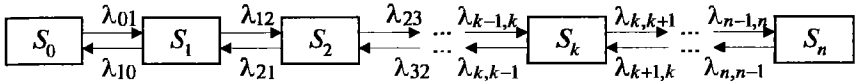


Рис. 7.7

Рассмотрим упорядоченное множество состояний системы S_0, S_1, S_2,..., S_k. Переходы могут осуществляться из любого состояния только в состояния с соседними номерами, т.е. из состояния S_k возможны переходы либо в состояние S_{k-1}, либо в состояние S_{k+1}[1].

Предположим, что все потоки событий, переводящие систему по стрелкам графа, простейшие с соответствующими интенсивностями $\lambda_{k,k+1}$ или $\lambda_{k+1,k}$.

По графу, представленному на рис. 7.7, составим и решим алгебраические уравнения для предельных вероятностей состояний (их существование вытекает из возможности перехода из каждого состояния в каждое другое и конечности числа состояний).

В соответствии с правилом составления таких уравнений (см. § 7.5) получим: для состояния S_0

$$\lambda_{01}p_0 = \lambda_{10}p_1, \qquad \textbf{(7.16)}$$

[1] При анализе численности популяций считают, что состояние S_k соответствует численности популяции, равной k, и переход системы из состояния S_k в состояние S_{k+1} происходит при рождении одного члена популяции, а переход в состояние S_{k-1} — при гибели одного члена популяции.

для состояния S_1 –

$$(\lambda_{12}+\lambda_{10})p_1 = \lambda_{01}p_0+\lambda_{21}p_2,$$

которое с учетом **(7.16)** приводится к виду:

$$\lambda_{12}p_1=\lambda_{21}p_2. \qquad (7.17)$$

Аналогично, записывая уравнения для предельных вероятностей других состояний, можно получить следующую систему уравнений:

$$\begin{cases} \lambda_{01}p_0 = \lambda_{10}p_1, \\ \lambda_{12}p_1 = \lambda_{21}p_2, \\ \ldots\ldots\ldots\ldots\ldots \\ \lambda_{k-1,k}p_{k-1} = \lambda_{k,k-1}p_k, \\ \ldots\ldots\ldots\ldots\ldots \\ \lambda_{n-1,n}p_{n-1} = \lambda_{n,n-1}p_n, \end{cases} \qquad (7.18)$$

к которой добавляется нормировочное условие

$$p_0 + p_1 + p_2 + \ldots + p_n = 1. \qquad (7.19)$$

Решая систему **(7.18)** и **(7.19)**, можно получить

$$p_0=\left(1+\frac{\lambda_{01}}{\lambda_{10}}+\frac{\lambda_{12}\lambda_{01}}{\lambda_{21}\lambda_{10}}+\ldots+\frac{\lambda_{n-1,n}\ldots\lambda_{12}\lambda_{01}}{\lambda_{n,n-1}\ldots\lambda_{21}\lambda_{10}}\right)^{-1}, \qquad (7.20)$$

$$p_1=\frac{\lambda_{01}}{\lambda_{10}}p_0,\ p_2=\frac{\lambda_{12}\lambda_{01}}{\lambda_{21}\lambda_{10}}p_0,\ldots,\ p_n=\frac{\lambda_{n-1,n}\ldots\lambda_{12}\lambda_{01}}{\lambda_{n,n-1}\ldots\lambda_{21}\lambda_{10}}p_0. \qquad (7.21)$$

Легко заметить, что в формулах **(7.21)** для p_1, p_2,..., p_n коэффициенты при p_0 — это слагаемые, стоящие после единицы в формуле **(7.20)**. Числители этих коэффициентов представляют собой произведения всех интенсивностей, стоящих у стрелок, ведущих слева направо от S_0 до данного состояния S_k ($k = 1, 2,..., n$), а знаменатели — произведения всех интенсивностей, стоящих у стрелок, ведущих справа налево из состояния S_k до S_0 (рис. 7.7).

▷ **Пример 7.6.** Процесс гибели и размножения представлен графом (рис.7.8). Найти предельные вероятности состояний.

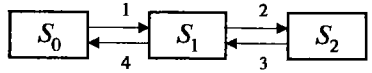


Рис. 7.8

Р е ш е н и е. По формуле **(7.20)** найдем

$$p_0=\left(1+\frac{1}{4}+\frac{2\cdot1}{3\cdot4}\right)^{-1}=0{,}706,$$

по **(7.21)**

$$p_1=\frac{1}{4}0{,}706=0{,}176,\ p_2=\frac{2\cdot1}{3\cdot4}0{,}706=0{,}118,$$

т.е. в установившемся стационарном режиме в среднем 70,6% времени система будет находиться в состоянии S_0, 17,6% — в состоянии S_1 и 11,8% — в состоянии S_2. ▶

7.7. СМО с отказами

В качестве показателей эффективности СМО с отказами будем рассматривать:

A — абсолютную пропускную способность СМО, т.е. среднее число заявок, обслуживаемых в единицу времени;

Q — относительную пропускную способность, т.е. среднюю долю пришедших заявок, обслуживаемых системой;

$P_{\text{отк}}$ — *вероятность отказа* — вероятность того, что заявка покинет СМО необслуженной;

k — среднее число занятых каналов (для многоканальной системы).

Одноканальная система с отказами. Рассмотрим задачу.

Имеется один канал, на который поступает поток заявок с интенсивностью λ. Поток обслуживаний имеет интенсивность μ[1]. Найти предельные вероятности состояний системы и показатели ее эффективности.

Система S (СМО) имеет два состояния: S_0 — канал свободен, S_1 — канал занят. Размеченный граф состояний представлен на рис. 7.9.

[1] Здесь и в дальнейшем предполагается, что все потоки событий, переводящие СМО из состояния в состояние, будут простейшими. К ним относится и поток обслуживаний — поток заявок, обслуживаемых одним непрерывно занятым каналом. Среднее время обслуживания $\bar{t}_{\text{об}}$ обратно по величине интенсивности μ, т.е. $\bar{t}_{\text{об}}=1/\mu$.

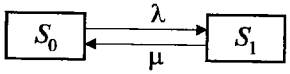


Рис. 7.9

В предельном стационарном режиме система алгебраических уравнений для вероятностей состояний **(7.14)** имеет вид (см. правило составления таких уравнений на с. 259):

$$\begin{cases} \lambda p_0 = \mu p_1, \\ \mu p_1 = \lambda p_0, \end{cases} \qquad (7.22)$$

т.е. система вырождается в одно уравнение. Учитывая нормировочное условие $p_0+p_1=1$, найдем из **(7.22)** предельные вероятности состояний

$$p_0 = \frac{\mu}{\lambda+\mu},\ p_1 = \frac{\lambda}{\lambda+\mu}, \qquad (7.23)$$

которые выражают среднее относительное время пребывания системы в состоянии S_0 (когда канал свободен) и S_1 (когда канал занят), т.е. определяют соответственно относительную пропускную способность Q системы и вероятность отказа $P_{\text{отк}}$:

$$Q = \frac{\mu}{\lambda+\mu}, \qquad (7.24)$$

$$P_{\text{отк}} = \frac{\lambda}{\lambda+\mu}. \qquad (7.25)$$

Абсолютную пропускную способность найдем, умножив относительную пропускную способность Q на интенсивность потока заявок λ:

$$A = \frac{\lambda\mu}{\lambda+\mu}. \qquad (7.26)$$

▷ **Пример 7.7.** Известно, что заявки на телефонные переговоры в телевизионном ателье поступают с интенсивностью λ, равной 90 заявок в час, а средняя продолжительность разговора по телефону $t_{\text{об}} = 2$ мин. Определить показатели эффективности работы СМО (телефонной связи) при наличии одного телефонного номера.

Р е ш е н и е. Имеем λ=90 (1/ч), $t_{\text{об}}$=2 мин. Интенсивность потока обслуживаний $\mu=1/t_{\text{об}}=1/2=0{,}5$ (1/мин)=30 (1/ч). По **(7.24)**

относительная пропускная способность СМО $Q = 30/(90+30) = 0{,}25$, т.е. в среднем только 25% поступающих заявок осуществят переговоры по телефону. Соответственно вероятность отказа в обслуживании составит $P_{\text{отк}}$=0,75 (см. **(7.25)**). Абсолютная пропускная способность СМО по **(7.26)** $A=90\cdot 0{,}25=22{,}5$, т.е. в среднем в час будут обслужены 22,5 заявки на переговоры. Очевидно, что при наличии только одного телефонного номера СМО будет плохо справляться с потоком заявок. ▶

Многоканальная система с отказами. Рассмотрим классическую *задачу Эрланга.*

Имеется n каналов, на которые поступает поток заявок с интенсивностью λ. Поток обслуживаний каждого канала имеет интенсивность μ. Найти предельные вероятности состояний системы и показатели ее эффективности.

Система S (СМО) имеет следующие состояния (нумеруем их по числу заявок, находящихся в системе): S_0, S_1, S_2,..., S_k,..., S_n, где S_k — состояние системы, когда в ней находится k заявок, т.е. занято k каналов.

Граф состояний СМО соответствует процессу гибели и размножения (рис. 7.10)

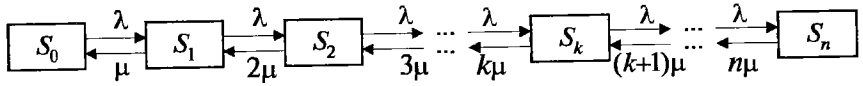


Рис. 7.10

Поток заявок последовательно переводит систему из любого левого состояния в соседнее правое с одной и той же интенсивностью λ. Интенсивность же потока обслуживаний, переводящих систему из любого правого состояния в соседнее левое, постоянно меняется в зависимости от состояния. Действительно, если СМО находится в состоянии S_2 (два канала заняты), то она может перейти в состояние S_1 (один канал занят), когда закончит обслуживание либо первый, либо второй канал, т.е. суммарная интенсивность их потоков обслуживаний будет 2μ. Аналогично суммарный поток обслуживаний, переводящий СМО из состояния S_3 (три канала заняты) в S_2, будет иметь интенсивность 3μ, т.е. может освободиться любой из трех каналов, и т.д.

В формуле **(7.20)** для схемы гибели и размножения получим для предельной вероятности состояния

$$p_0 = \left(1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2!\mu^2} + \ldots + \frac{\lambda^k}{k!\mu^k} + \ldots + \frac{\lambda^n}{n!\mu^n}\right)^{-1}, \qquad (7.27)$$

где члены разложения $\frac{\lambda}{\mu}, \frac{\lambda^2}{2!\mu^2}, \ldots, \frac{\lambda^n}{n!\mu^n}$ — коэффициенты при p_0 в выражениях для предельных вероятностей $p_1, p_2, \ldots, p_k, \ldots, p_n$. Величина

$$\rho = \frac{\lambda}{\mu} \qquad (7.28)$$

называется *приведенной интенсивностью потока заявок* или *интенсивностью нагрузки канала*. Она выражает среднее число заявок, приходящих за среднее время обслуживания одной заявки. Теперь

$$p_0 = \left(1 + \rho + \frac{\rho^2}{2!} + \ldots + \frac{\rho^k}{k!} + \ldots + \frac{\rho^n}{n!}\right)^{-1}, \qquad (7.29)$$

$$p_1 = \rho p_0,\ p_2 = \frac{\rho^2}{2!} p_0, \ldots,\ p_k = \frac{\rho^k}{k!} p_0, \ldots,\ p_n = \frac{\rho^n}{n!} p_0. \qquad (7.30)$$

Формулы (7.29) и (7.30) для предельных вероятностей получили названия *формул Эрланга* в честь основателя теории массового обслуживания.

Вероятность отказа СМО есть предельная вероятность того, что все n каналов системы будут заняты, т.е.

$$P_{\text{отк}} = \frac{\rho^n}{n!} p_0. \qquad (7.31)$$

Относительная пропускная способность — вероятность того, что заявка будет обслужена:

$$Q = 1 - P_{\text{отк}} = 1 - \frac{\rho^n}{n!} p_0. \qquad (7.32)$$

Абсолютная пропускная способность:

$$A = \lambda Q = \lambda\left(1 - \frac{\rho^n}{n!} p_0\right). \qquad (7.33)$$

Среднее число занятых каналов, т.е. математическое ожидание числа занятых каналов:

$$\bar{k} = \sum_{k=0}^{n} k p_k,$$

где p_k — предельные вероятности состояний, определяемые по формулам (7.29), (7.30).

Однако среднее число занятых каналов можно найти проще, если учесть, что абсолютная пропускная способность системы A есть не что иное, как интенсивность потока *обслуженных* системой заявок (в единицу времени). Так как каждый занятый канал обслуживает в среднем μ заявок (в единицу времени), то среднее число занятых каналов

$$\bar{k} = \frac{A}{\mu} \qquad (7.34)$$

или, учитывая (7.33), (7.28),

$$\bar{k} = \rho\left(1 - \frac{\rho^n}{n!} p_0\right). \qquad (7.35)$$

▷ **Пример 7.8.** В условиях примера 7.7 определить оптимальное число телефонных номеров в телевизионном ателье, если условием оптимальности считать удовлетворение из каждых 100 заявок на переговоры в среднем не менее 90 заявок.

Р е ш е н и е. Интенсивность нагрузки канала по формуле (7.28) $\rho = 90/30 = 3$, т.е. за время среднего (по продолжительности) телефонного разговора $t_{\text{об}} = 2$ мин поступает в среднем 3 заявки на переговоры.

Будем постепенно увеличивать число каналов (телефонных номеров) $n = 2, 3, 4, \ldots$ и определим по формулам (7.29), (7.32), (7.33) для получаемой n-канальной СМО характеристики обслуживания. Например, при $n = 2$ $p_0 = \left(1 + 3 + 3^2/2!\right)^{-1} = 0{,}118 \approx 0{,}12$; $Q = 1 - \left(3^2/2!\right) \cdot 0{,}118 \approx 0{,}471$; $A = 90 \cdot 0{,}471 = 42{,}4$. Значение характеристик СМО сведем в табл. 7.1.

Таблица 7.1

Характеристика обслуживания	*Обозначение*	*Число каналов (телефонных номеров)*					
		1	2	3	4	5	6
Относительная пропускная способность	Q	0,25	0,47	0,65	0,79	0,90	0,95
Абсолютная пропускная способность	A	22,5	42,4	58,8	71,5	80,1	85,3

По условию оптимальности $Q \geq 0{,}9$, следовательно, в телевизионном ателье необходимо установить 5 телефонных номеров (в этом случае $Q = 0{,}90$ — см. табл. 7.1). При этом в час будут обслуживаться в среднем 80 заявок ($A = 80{,}1$), а среднее число занятых телефонных номеров (каналов) по формуле (**7.34**) $\overline{k} = 80{,}1/30 = 2{,}67$. ▶

▷ **Пример 7.9.** В вычислительный центр коллективного пользования с тремя ЭВМ поступают заказы от предприятий на вычислительные работы. Если работают все три ЭВМ, то вновь поступающий заказ не принимается, и предприятие вынуждено обратиться в другой вычислительный центр. Среднее время работы с одним заказом составляет 3 ч. Интенсивность потока заявок 0,25 1/ч. Найти предельные вероятности состояний и показатели эффективности работы вычислительного центра.

Р е ш е н и е. По условию n=3, λ=0,25 1/ч, $t_{об}$=3 ч. Интенсивность потока обслуживаний $\mu = 1/t_{об} = 1/3 = 0{,}33$. Интенсивность нагрузки ЭВМ по формуле (**7.28**) $\rho = 0{,}25/0{,}33 = 0{,}75$. Найдем предельные вероятности состояний:

по формуле (**7.29**): $p_0 = (1+0{,}75+0{,}75^2/2!+0{,}75^3/3!)^{-1} = 0{,}476$;

по формуле (**7.30**): $p_1 = 0{,}75 \cdot 0{,}476 = 0{,}357$; $p_2 = (0{,}75^2/2!) \cdot 0{,}476 = 0{,}134$; $p_3 = (0{,}75^3/3!) \cdot 0{,}476 = 0{,}033$, т.е. в стационарном режиме работы вычислительного центра в среднем 47,6% времени нет ни одной заявки, 35,7% — имеется одна заявка (занята одна ЭВМ), 13,4% — две заявки (две ЭВМ), 3,3% времени — три заявки (заняты три ЭВМ).

Вероятность отказа (когда заняты все три ЭВМ), таким образом, $P_{отк} = p_3 = 0{,}033$.

Согласно формуле (**7.32**) относительная пропускная способность центра $Q = 1 - 0{,}033 = 0{,}967$, т.е. в среднем из каждых 100 заявок вычислительный центр обслуживает 96,7 заявок.

По формуле (**7.33**) абсолютная пропускная способность центра $A = 0{,}25 \cdot 0{,}967 = 0{,}242$, т.е. в один час в среднем обслуживается 0,242 заявки.

Согласно формуле (**7.34**) среднее число занятых ЭВМ $\overline{k} = 0{,}242/0{,}33 = 0{,}725$, т.е. каждая из трех ЭВМ будет занята обслуживанием заявок в среднем лишь на $72{,}5/3 = 24{,}2\%$.

При оценке эффективности работы вычислительного центра необходимо сопоставить доходы от выполнения заявок с потерями от простоя дорогостоящих ЭВМ (с одной стороны, здесь высокая пропускная способность СМО, а с другой — значительный простой каналов обслуживания) и выбрать компромиссное решение. ▶

7.8. Понятие о методе статистических испытаний (методе Монте-Карло)

Основное допущение, при котором анализировались рассмотренные выше СМО, состоит в том, что все потоки событий, переводящие их из состояния в состояние, были простейшими. При нарушении этого требования общих аналитических методов для таких систем не существует. Имеются лишь отдельные результаты, позволяющие выразить в аналитическом виде характеристики СМО через параметры задачи.

В случае, когда для анализа работы СМО аналитические методы неприменимы (или же требуется проверить их точность), используют универсальный *метод статистических испытаний*, или, как его называют, *метод Монте-Карло*. Название «метод Монте-Карло» связано с тем, что одним из возможных способов имитации случайных явлений является рулетка, игрой в которую знаменит город Монте-Карло.

Для уяснения сути метода Монте-Карло рассмотрим задачу о вычислении определенного интеграла.

Пусть требуется вычислить определенный интеграл $\int_0^1 f(x)dx$, где $f(x)$ — непрерывная на отрезке [0; 1] элементарная функция, причем $0 \leq f(x) \leq 1$. Задача носит достаточно общий характер, так как к данному интегралу путем соответствующих линейных подстановок (замен) может быть сведен любой интеграл $\int_a^b f(x)dx$.

Геометрически интеграл $\int_0^1 f(x)dx$ представляет собой площадь области S под кривой $y = f(x)$ (рис. 7.11).

Будем многократно повторять опыт, состоящий в бросании наудачу точки в единичный квадрат.

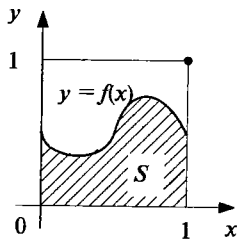


Рис. 7.11

Используя геометрическое определение вероятности (§ 1.4), вероятность события A — попадания брошенной случайной точки в область S — есть отношение площади области S к площади квадрата, т.е.

$$p(A) = \frac{S}{1^2} = \int_0^1 f(x)dx.$$

В соответствии с теоремой Бернулли при достаточно большом числе опытов n в качестве оценки искомой вероятности $p(A)$ следует взять частость m/n, которая и будет являться приближенным значением интеграла по результатам случайных испытаний.

Отличительная особенность применения метода Монте-Карло состоит в том, что границы, в которых будет заключено истинное значение неизвестной величины (в данной задаче — вычисляемого интеграла), мы можем указать не точно, а лишь с некоторой вероятностью, называемой *доверительной* (об этом речь пойдет в § 9.6).

Конечно, при реализации метода Монте-Карло не требуется проведения реальных опытов по бросанию случайной точки в квадрат (или, например в других задачах, по бросанию монеты, игральной кости и т.п.). Для этой цели служат специальные компьютерные программы или датчики случайных чисел (точнее «псевдослучайных» чисел, ибо последние не в полной мере удовлетворяют требованию случайности).

Применение метода Монте-Карло в системах массового обслуживания состоит в том, что вместо аналитического описания СМО проводится «розыгрыш» случайного процесса, проходящего в СМО, с помощью специально организованной процедуры. В результате такого «розыгрыша» получается каждый раз новая, отличная от других реализация случайного процесса. (Реализации случайного процесса называются *статистическими испытаниями* — отсюда и второе название метода.) Это множество реализаций можно использовать как некий искусственно полученный статистический материал, который обрабатывается обычными методами математической статистики. После такой обработки могут быть получены приближенно любые характеристики обслуживания.

При моделировании случайных явлений методом Монте-Карло мы пользуемся самой случайностью как аппаратом исследования. Для сложных систем обслуживания с немарковским случайным

процессом метод статистических испытаний, как правило, оказывается проще аналитического, а часто и вовсе единственно возможным.

В данной главе мы ограничились рассмотрением СМО с отказами. С другими системами массового обслуживания, например СМО с ожиданием (одноканальными и многоканальными, с ограниченным и неограниченным временем ожидания и др.), можно ознакомиться, например, в пособии [15].

Упражнения

7.10. Случайный процесс определяется формулой $X(t) = Xe^{-t}(t > 0)$, где X — случайная величина, распределенная по нормальному закону с параметрами a и σ^2. Найти математическое ожидание, дисперсию, корреляционную и нормированную корреляционную функции случайного процесса.

7.11. Построить граф состояний следующего случайного процесса: система состоит из двух автоматов по продаже газированной воды, каждый из которых в случайный момент времени может быть либо занятым, либо свободным.

7.12. Построить граф состояний системы S, представляющей собой электрическую лампочку, которая в случайный момент времени может быть либо включена, либо выключена, либо выведена из строя.

7.13. Среднее число заказов на такси, поступающих на диспетчерский пункт в одну минуту, равно 3. Найти вероятность того, что за две минуты поступит: а) 4 вызова; б) хотя бы один; в) ни одного вызова. (Поток заявок простейший.)

7.14. Найти предельные вероятности для систем S, граф которых изображен на рис. 7.11 и 7.12.

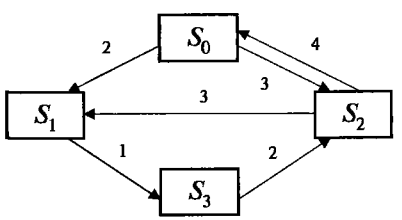


Рис. 7.11

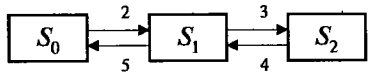


Рис. 7.12

7.15. Рассматривается круглосуточная работа пункта проведения профилактического осмотра автомашин с одним каналом (одной группой проведения осмотра). На ос-

мотр и выявление дефектов каждой машины затрачивается в среднем 0,5 ч. На осмотр поступает в среднем 36 машин в сутки. Потоки заявок и обслуживаний — простейшие. Если машина, прибывшая в пункт осмотра, не застает ни одного канала свободным, она покидает пункт осмотра необслуженной. Определить предельные вероятности состояний и характеристики обслуживания профилактического пункта осмотра.

7.16. Решить задачу **7.15** для случая $n = 4$ канала (групп проведения осмотра). Найти минимальное число каналов, при котором относительная пропускная способность пункта осмотра будет не менее 0,9.

7.17. Одноканальная СМО с отказами представляет собой одну телефонную линию, на вход которой поступает простейший поток вызовов с интенсивностью 0,4 вызовов/мин. Средняя продолжительность разговора 3 мин.; время разговора имеет показательное распределение. Найти предельные вероятности состояний и характеристики обслуживания СМО. Сравнить пропускную способность СМО с номинальной, которая была бы, если разговор длился в точности 3 мин., а заявки шли одна за другой регулярно, без перерывов.

7.18. Имеется двухканальная простейшая СМО с отказами. На ее вход поступает поток заявок с интенсивностью 4 заявки/ч. Среднее время обслуживания одной заявки 0,8 ч. Каждая обслуженная заявка приносит доход 4 ден. ед. Содержание каждого канала обходится 2 ден. ед./ч. Выяснить, выгодно или невыгодно в экономическом отношении увеличить число каналов до трех.

Раздел II
Математическая статистика

Глава 8 Вариационные ряды и их характеристики

8.1. Вариационные ряды и их графическое изображение

Установление статистических закономерностей, присущих массовым случайным явлениям, основано на изучении статистических данных — сведений о том, какие значения принял в результате наблюдений интересующий нас признак (случайная величина X).

▷ **Пример 8.1.** Необходимо изучить изменение выработки на одного рабочего механического цеха в отчетном году по сравнению с предыдущим. Получены следующие данные о распределении 100 рабочих цеха по выработке в отчетном году (в процентах к предыдущему году):

$$\underbrace{97{,}8;\ 97{,}0;\ 101{,}7;\ 132{,}5;\ \ldots;\ 142{,}3;\ 104{,}2;\ 141{,}0;\ 122{,}1}_{100\text{ значений}}.$$

Различные значения признака (случайной величины X) называются *вариантами* (обозначаем их через x).

Рассмотрение и осмысление этих данных (особенно при большом числе наблюдений n) затруднительно, и по ним практически нельзя представить характер распределения признака (случайной величины X).

Первый шаг к осмыслению имеющегося статистического материала — это его упорядочение, расположение вариантов в порядке возрастания (убывания), т.е. *ранжирование* вариантов ряда:

$$x_{\min} = \underbrace{94{,}0;\ 94{,}2;\ \ldots;\ 142{,}3;\ 141{,}0}_{n=100\text{ значений}} = x_{\max}.$$

В таком виде изучать выработку рабочих тоже не очень удобно из-за обилия числовых данных. Поэтому разобьем варианты на отдельные интервалы, т.е. проведем их *группировку*.

Число интервалов m следует брать не очень большим, чтобы после группировки ряд не был громоздким, и не очень малым, чтобы не потерять особенности распределения признака.

Согласно *формуле Стерджеса* рекомендуемое число интервалов $m = 1 + 3{,}322\lg n$, а *величина интервала* (*интервальная разность, ширина интервала*)

$$k = \frac{x_{\max} - x_{\min}}{1 + 3{,}322\lg n},$$

где $x_{\max} - x_{\min}$ — разность между наибольшим и наименьшим значениями признака.

В примере **8.1** $k = (141{,}0 — 97{,}0)/(1+3{,}322\lg 100) = 5{,}76(\%)$.

Примем $k = 6{,}0(\%)$. За начало первого интервала рекомендуется брать величину $x_{\text{нач}} = x_{\min} - k/2$. В данном случае $x_{\text{нач}} = 97{,}0 - 6{,}0/2 = 94{,}0(\%)$.

Сгруппированный ряд представим в виде таблицы.

Таблица 8.1

i	*Выработка в отчетном году в процентах к предыдущему* x	*Частота (количество рабочих)* n_i	*Частость (доля рабочих)* $w_i = \frac{n_i}{n}$	*Накопленная частота* $n_i^{\text{нак}}$	*Накопленная частость* $w_i^{\text{нак}} = \frac{n_i^{\text{нак}}}{n}$
1	94,0—100,0	3	0,03	3	0,03
2	100,0—106,0	7	0,07	10	0,10
3	106,0—112,0	11	0,11	21	0,21
4	112,0—118,0	20	0,20	41	0,41
5	118,0—124,0	28	0,28	69	0,69
6	124,0—130,0	19	0,19	88	0,88
7	130,0—136,0	10	0,10	98	0,98
8	136,0—142,0	2	0,02	100	1,00
	$\sum$	100	1,00	—	—

Числа, показывающие, сколько раз встречаются варианты из данного интервала, называются *частотами* (обозначаем n_i), а отношение их к общему числу наблюдений — *частостями* или *относительными частотами*, т.е. $w_i = n_i/n$. Частоты и частости называются *весами*.

Определение. ***Вариационным рядом*** *называется ранжированный в порядке возрастания (или убывания) ряд вариантов с соответствующими им весами (частотами или частостями)*[1].

[1] Если вариант x_i (i = 1, 2, ..., n) вариационного ряда рассматривается как случайная величина X_i, получаемая в результате многократного наблюдения интересующего нас признака X, то X_i называется *порядковой статистикой*.

Если просмотр первичных, несгруппированных данных делал затруднительным представление об изменчивости значений признака, то полученный теперь вариационный ряд позволяет выявить закономерности распределения рабочих по интервалам выработки. Мы видим, например, что выработка колеблется от 94,0 до 142,0%, наибольшее число рабочих (28, или 0,28 от общего числа) увеличили выработку до 118,0—124,0%, уменьшили выработку (в пределах от 94,0 до 100%) 3 рабочих и т.п.

При изучении вариационных рядов наряду с понятием частоты используется понятие *накопленной частоты* (обозначаем $n_i^{\text{нак}}$). Накопленная частота показывает, сколько наблюдалось вариантов со значением признака, меньшим x. Отношение накопленной частоты $n_i^{\text{нак}}$ к общему числу наблюдений n назовем *накопленной частостью* $w_i^{\text{нак}}$.

Накопленные частоты (частости) для каждого интервала находятся последовательным суммированием частот (частостей) всех предшествующих интервалов, включая данный (см. табл. 8.1). Например, для $x = 124$ накопленная частота $n_i^{\text{нак}} = 3 + 7 + 11 + 20 + 28 = 69$, т.е. 69 рабочих имели выработку, меньшую 124%.

Для задания вариационного ряда достаточно указать варианты и соответствующие им частоты (частости) или накопленные частоты (частости) (в табл. 8.1 приведены и те, и другие).

Вариационный ряд называется *дискретным*, если любые его варианты отличаются на постоянную величину, и — *непрерывным* (*интервальным*), если варианты могут отличаться один от другого на сколь угодно малую величину. Так, вариационный ряд, представленный в табл. 8.1, — интервальный (проценты выработки условно округлены до десятых долей). Примером дискретного ряда является распределение 50 рабочих механического цеха по тарифному разряду (табл. 8.2).

Таблица 8.2

Тарифный разряд x_i	1	2	3	4	5	6	Σ
Частота (количество рабочих) n_i	2	3	6	8	22	9	50

Для графического изображения вариационных рядов наиболее часто используются полигон, гистограмма, кумулятивная кривая.

Полигон, как правило, служит для изображения дискретного вариационного ряда и представляет собой ломаную, в которой концы отрезков прямой имеют координаты (x_i, n_i), $i = 1, 2, ..., m$.

Гистограмма служит только для изображения интервальных вариационных рядов и представляет собой ступенчатую фигуру из прямоугольников с основаниями, равными интервалам значений признака $k_i = x_{i+1} - x_i$, $i = 1, 2, ..., m$, и высотами, равными частотам (частостям) n_i (w_i) интервалов. Если соединить середины верхних оснований прямоугольников отрезками прямой, то можно получить полигон того же распределения.

Кумулятивная кривая (*кумулята*) — кривая накопленных частот (частостей). Для дискретного ряда кумулята представляет ломаную, соединяющую точки $(x_i, n_i^{\text{нак}})$ или $(x_i, w_i^{\text{нак}})$, $i = 1, 2, ..., m$. Для интервального вариационного ряда ломаная начинается с точки, абсцисса которой равна началу первого интервала, а ордината — накопленной частоте (частости), равной нулю. Другие точки этой ломаной соответствуют концам интервалов.

Весьма важным является понятие эмпирической функции распределения.

О п р е д е л е н и е. ***Эмпирической функцией распределения*** $F_n(x)$ *называется относительная частота* (*частость*) *того, что признак* (*случайная величина* X) *примет значение, меньшее заданного* x, т.е.

$$F_n(x) = w(X < x) = w_x^{\text{нак}}. \qquad \textbf{(8.1)}$$

Другими словами, для данного x эмпирическая функция распределения представляет накопленную частость $w_x^{\text{нак}} = n_x^{\text{нак}}/n$.

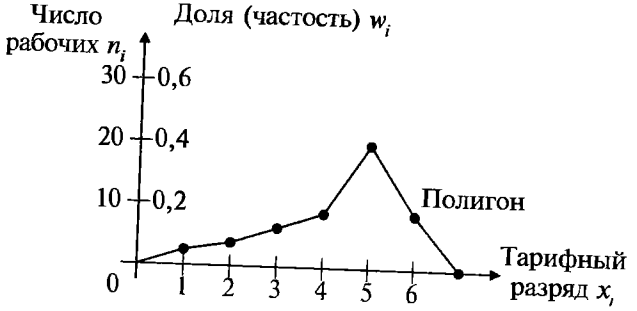
▷ **Пример 8.2.** Построить полигон (гистограмму), кумуляту и эмпирическую функцию распределения рабочих:

а) по тарифному разряду по данным табл. 8.2;

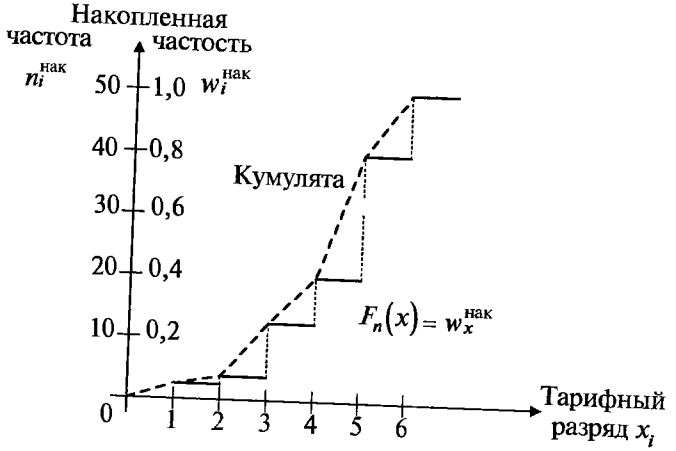
б) по выработке по данным табл. 8.1.

Р е ш е н и е. На рис. 8.1 и 8.2 изображены полигон (гистограмма), кумулята и эмпирическая функция распределения соответственно для дискретного (табл. 8.2) и интервального (табл. 8.1) вариационных рядов. Обращаем внимание на то, что для дискретного вариационного ряда эмпирическая функция распределения представляет собой разрывную ступенчатую функцию по аналогии с функцией распределения для дискретной случайной величины (§ 3.5) с той лишь разницей,

что теперь по оси ординат вместо вероятностей располагаются частости (см. рис 8.1).



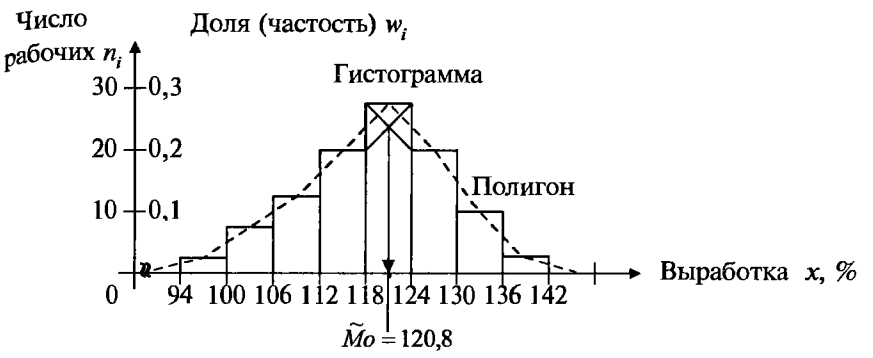
а)



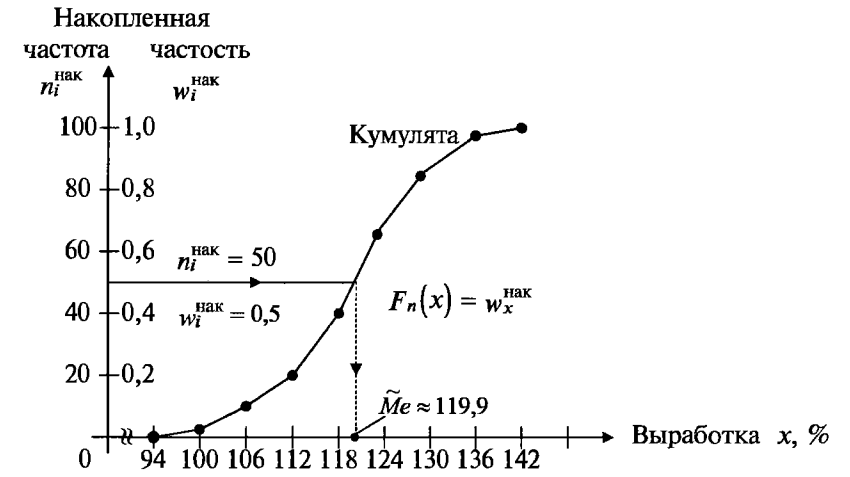
б)

Рис. 8.1

Для интервального вариационного ряда (табл. 8.1) имеем лишь значения функции распределения $F_n(x)$ на концах интервала (см. последнюю графу табл. 8.1). Поэтому для графического изображения этой функции целесообразно ее доопределить, соединив точки графика, соответствующие концам интервалов, отрезками прямой. В результате полученная ломаная совпадет с кумулятой (см. рис. 8.2*б*). ▶



а)



б)

Рис. 8.2

Вариационный ряд является статистическим аналогом (реализацией) распределения признака (случайной величины X). В этом смысле полигон (гистограмма) аналогичен кривой распределения, а эмпирическая функция распределения — функции распределения случайной величины *X*.

Вариационный ряд содержит достаточно полную информацию об изменчивости (вариации) признака. Однако обилие числовых данных, с помощью которых он задается, усложняет их использование. В то же время на практике часто оказывается достаточным знание лишь сводных характеристик вариационных рядов: средних или характеристик центральной тенденции; характеристик изменчивости (вариации) и др. Расчет статисти-

ческих характеристик представляет собой второй после группировки этап обработки данных наблюдений.

8.2. Средние величины

Средние величины характеризуют значения признака, вокруг которого концентрируются наблюдения или, как говорят, центральную тенденцию распределения. Наиболее распространенной из средних величин является средняя арифметическая.

Определение. ***Средней арифметической*** *вариационного ряда называется сумма произведений всех вариантов на соответствующие частоты, деленная на сумму частот*:

$$\bar{x} = \frac{\sum_{i=1}^{m} x_i n_i}{n}, \qquad (8.2)$$

где x_i — варианты дискретного ряда или середины интервалов интервального вариационного ряда; n_i — соответствующие им частоты; m — число неповторяющихся вариантов или число интервалов; $n = \sum_{i=1}^{m} n_i$.

Очевидно, что

$$\bar{x} = \sum_{i=1}^{m} x_i w_i,$$

где $w_i = n_i/n$ — частости вариантов или интервалов.

▷ **Пример 8.3.** Найти среднюю выработку рабочих по данным табл. 8.1.

Решение. По формуле **(8.2)** для интервального вариационного ряда

$\bar{x} = \dfrac{97 \cdot 3 + 103 \cdot 7 + \ldots + 133 \cdot 10 + 139 \cdot 2}{100} = 119{,}2(\%)$, где числа 97, 103,..., 133,139 — середины соответствующих интервалов. ▶

Для несгруппированного ряда все частоты n_i=1 (i=1,2,..., n), а

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (8.3)$$

есть «невзвешенная» средняя арифметическая.

Отметим **основные свойства средней арифметической**, аналогичные свойствам математического ожидания случайной величины:

1. *Средняя арифметическая постоянной равна самой постоянной.*

2. *Если все варианты увеличить (уменьшить) в одно и то же число раз, то средняя арифметическая увеличится (уменьшится) во столько же раз*:

$$\overline{kx} = k\bar{x} \text{ или } \frac{\sum_{i=1}^{m} (kx_i) n_i}{n} = k \frac{\sum_{i=1}^{m} x_i n_i}{n}.$$

3. *Если все варианты увеличить (уменьшить) на одно и то же число, то средняя арифметическая увеличится (уменьшится) на то же число*:

$$\overline{x + c} = \bar{x} + c \text{ или } \frac{\sum_{i=1}^{m} (x_i + c) n_i}{n} = \frac{\sum_{i=1}^{m} x_i n_i}{n} + c.$$

4. *Средняя арифметическая отклонений вариантов от средней арифметической равна нулю*:

$$\overline{x - \bar{x}} = 0 \text{ или } \sum_{i=1}^{m} (x_i - \bar{x}) n_i = 0. \qquad (8.4)$$

□ При $c = \bar{x}$ $\overline{x - c} = \bar{x} - c = \bar{x} - \bar{x} = 0$. ■

5. *Средняя арифметическая алгебраической суммы нескольких признаков равна такой же сумме средних арифметических этих признаков*:

$$\overline{x + y} = \bar{x} + \bar{y}.$$

6. *Если ряд состоит из нескольких групп, общая средняя равна средней арифметической групповых средних, причем весами являются объемы групп*:

$$\bar{x} = \frac{\sum_{i=1}^{l} \bar{x}_i n_i}{n}, \qquad (8.5)$$

где $\bar{x}$ — общая средняя (средняя арифметическая всего ряда);

$\bar{x}_i$ — групповая средняя i-й группы, объем которой равен n_i;

l — число групп.

При решении практических задач могут применяться и иные формы средней, которые можно получить из *средней степенной k-го порядка*[1]:

$$\bar{x}_k = \sqrt[k]{\frac{\sum_{i=1}^{m} x_i^k n_i}{n}}, \text{ где } x_i > 0.$$

Легко убедиться в том, что при $k = 1$ получаем формулу *средней арифметической*. При других значениях k получаем формулы:

$k=-1$ $\bar{x}_{-1} = \left(\frac{\sum_{i=1}^{m} x^{-1} n_i}{n} \right)^{-1} = \frac{n}{\sum_{i=1}^{m} \frac{n_i}{x_i}}$ — *средней гармонической*;

$k = 0$ (после раскрытия неопределенности при вычислении предела $\lim\limits_{k \to 0} \bar{x}_k$) $\bar{x}_0 = \sqrt[n]{x_1^{n_1} x_2^{n_2} \ldots x_m^{n_m}} = \sqrt[n]{\prod_{i=1}^{m} x_i^{n_i}}$ — *средней геометрической*;

$k = 2$ $\bar{x}_2 = \sqrt{\frac{\sum_{i=1}^{m} x_i^2 n_i}{n}}$ — *средней квадратической* и т.д.

Можно показать, что с ростом порядка k степенная средняя возрастает, т.е. $\bar{x}_{-1} < \bar{x}_0 < \bar{x}_1 < \bar{x}_2 < \ldots$ (свойство *мажорантности средних*).

Кроме рассмотренных средних величин, называемых *аналитическими*, в статистическом анализе применяют *структурные*, или *порядковые*, средние. Из них наиболее широко применяются медиана и мода.

Определение. ***Медианой $\tilde{Me}$ вариационного ряда*** *называется значение признака, приходящееся на середину ранжированного ряда наблюдений.*

---

[1] Более корректна запись: $\bar{x}_k = \frac{\left(\sum_{i=1}^{m} x_i^k n_i \right)^{\frac{1}{k}}}{n}$, так как корень k-й степени определяется только для натуральных $k \geq 2$.

Для дискретного вариационного ряда с нечетным числом членов медиана равна серединному варианту, а для ряда с четным числом членов — полусумме двух серединных вариантов.

▷ **Пример 8.4.** Найти медиану распределения рабочих по тарифному разряду по данным табл. 8.2.

Решение. $n = 50$ — четное, следовательно, серединных вариантов два: $x_{25} = 5$ и $x_{26} = 5$. Поэтому $\tilde{Me} = (x_{25}+x_{26})/2 = (5+5)/2 = 5\%$). ▶

Для интервального вариационного ряда находится медианный интервал, на который приходится середина ряда, а значение медианы на этом интервале находят с помощью линейного интерполирования. Не приводя соответствующей формулы, отметим, что медиана может быть приближенно найдена с помощью кумуляты как значение признака, для которого $n_x^{\text{нак}} = n/2$ или $w_x^{\text{нак}} = 1/2$.

Достоинство медианы как меры центральной тенденции заключается в том, что на нее не влияет изменение крайних членов вариационного ряда, если любой из них, меньший медианы, остается меньше ее, а любой, больший медианы, продолжает быть больше ее. Медиана предпочтительнее средней арифметической для ряда, у которого крайние варианты по сравнению с остальными оказались чрезмерно большими или малыми.

Определение. ***Модой $\tilde{Mo}$ вариационного ряда*** *называется вариант, которому соответствует наибольшая частота.*

Например, для вариационного ряда табл. 8.2 мода $\tilde{Mo} = 5$, так как этому варианту соответствует наибольшая частота $n_i = 22$. Для интервального ряда находится модальный интервал, имеющий наибольшую частоту, а значение моды на этом интервале определяют с помощью линейного интерполирования. Однако проще моду можно найти графическим путем с помощью гистограммы.

Особенность моды как меры центральной тенденции заключается в том, что она не изменяется при изменении крайних членов ряда, т.е. обладает определенной устойчивостью к вариации признака.

▷ **Пример 8.5.** Найти медиану и моду распределения рабочих по выработке по данным табл. 8.1.

Р е ш е н и е. На рис. 8.2*б* проведем горизонтальную прямую y=0,5 (или y=50), соответствующую накопленной частости $w_x^{\text{нак}} = F_n(x) = 0{,}5$ (или накопленной частоте $n_x^{\text{нак}} = 50$), до пересечения с графиком эмпирической функции распределения (или кумулятой). Абсцисса точки пересечения и будет медианой вариационного ряда: $\tilde{M}e = 119{,}9(\%)$.

На гистограмме распределения (рис. 8.2*а*) находим прямоугольник с наибольшей частотой (частостью). Соединяя отрезками прямых вершины этого прямоугольника с соответствующими вершинами двух соседних прямоугольников (см. рис. 8.2*а*), получим точку пересечения этих отрезков (диагоналей), абсцисса которой и будет модой вариационного ряда: $\tilde{M}o = 120{,}8(\%)$. ▶

8.3. Показатели вариации

Средние величины, рассмотренные выше, не отражают изменчивости (вариации) значений признака.

Простейшим (и весьма приближенным) показателем вариации является *вариационный размах* R, равный разности между наибольшим и наименьшим вариантами ряда:

$$R = x_{\max} - x_{\min}.$$

Наибольший интерес представляют меры вариации (рассеяния) наблюдений вокруг средних величин, в частности, вокруг средней арифметической.

Средним линейным отклонением вариационного ряда называется средняя арифметическая абсолютных величин отклонений вариантов от их средней арифметической:

$$d = \frac{\sum_{i=1}^{m} |x_i - \bar{x}| n_i}{n}. \qquad (8.6)$$

(Заметим, что «простая» сумма отклонений $\sum_{i=1}^{m} (x_i - \bar{x}) n_i$ не может характеризовать вариацию признака, ибо согласно свойству **4** средней арифметической эта сумма равна нулю для любого вариационного ряда.)

О п р е д е л е н и е. ***Дисперсией s^2 вариационного ряда*** *называется средняя арифметическая квадратов отклонений вариантов от их средней арифметической:*

$$s^2 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 n_i}{n}. \qquad (8.7)$$

Формулу для дисперсии вариационного ряда можно записать в виде:

$$s^2 = \sum_{i=1}^{m} (x_i - \bar{x})^2 w_i,$$

где $w_i = n_i / n$.

Для несгруппированного ряда ($n_i = 1$) по формуле **(8.7)** имеем:

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

Дисперсию s^2 часто называют *эмпирической* или *выборочной*, подчеркивая, что она (в отличие от дисперсии случайной величины σ^2) находится по опытным или стати тическим данным.

Жёлательно в качестве меры вариации (рассеяния) иметь характеристику, выраженную в тех же единицах, что и значения признака. Такой характеристикой является ***среднее квадратическое отклонение*** s *— арифметическое значение корня квадратного из дисперсии —*

$$s = \sqrt{\frac{\sum_{i=1}^{m} (x_i - x)^2 n_i}{n}}. \qquad (8.8)$$

Рассматривается также безразмерная характеристика — *коэффициент вариации*, равный процентному отношению среднего квадратического отклонения к средней арифметической:

$$\tilde{v} = \frac{s}{\bar{x}} \cdot 100\% \quad (\bar{x} \neq 0). \qquad (8.9)$$

Если коэффициент вариации признака, принимающего только положительные значения, высок (например, более 100%), то, как правило, это свидетельствует о неоднородности значений признака.

Отметим **основные свойства дисперсии**, аналогичные свойствам дисперсии случайной величины:

1. *Дисперсия постоянной равна нулю.*

2. *Если все варианты увеличить (уменьшить) в одно и то же число k раз, то дисперсия увеличится (уменьшится) в k^2 раз:*

$$s_{kx}^2 = k^2 s_x^2 \text{ или } \frac{\sum_{i=1}^{m}(x_i k - \bar{x}k)^2 n_i}{n} = k^2 \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2 n_i}{n}.$$

3. *Если все варианты увеличить (уменьшить) на одно и то же число, то дисперсия не изменится:*

$$s_{x+c}^2 = s_x^2 = s^2 \text{ или } \frac{\sum_{i=1}^{m}[(x_i + c) - (\bar{x} + c)]^2 n_i}{n} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2 n_i}{n}.$$

4. *Дисперсия равна разности между средней арифметической квадратов вариантов и квадратом средней арифметической:*

$$s^2 = \overline{x^2} - \bar{x}^2, \qquad \textbf{(8.10)}$$

где

$$\overline{x^2} = \frac{\sum_{i=1}^{m} x_i^2 n_i}{n}. \qquad \textbf{(8.11)}$$

□ $$s^2 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2 n_i}{n} = \frac{\sum_{i=1}^{m} x_i^2 n_i}{n} - 2\bar{x}\frac{\sum_{i=1}^{m} x_i n_i}{n} + \bar{x}^2 \frac{\sum_{i=1}^{m} n_i}{n} =$$

$$= \overline{x^2} - 2\bar{x}\cdot\bar{x} + \bar{x}^2 = \overline{x^2} - \bar{x}^2, \text{ ибо } \sum_{i=1}^{m} n_i = n. \blacksquare$$

5. *Если ряд состоит из нескольких групп наблюдений, то общая дисперсия равна сумме средней арифметической групповых дисперсий и межгрупповой дисперсии:*

$$s^2 = \overline{s_i^2} + \delta^2, \qquad \textbf{(8.12)}$$

где s^2 — общая дисперсия (дисперсия всего ряда);

$$\overline{s_i^2} = \frac{\sum_{i=1}^{l} s_i^2 n_i}{n} \qquad \textbf{(8.13)}$$

— средняя арифметическая групповых дисперсий;

$$s_i^2 = \frac{\sum_{j=1}^{m}(x_j - \bar{x}_i)^2 n_j}{n_i}; \qquad \textbf{(8.14)}$$

$$\delta^2 = \frac{\sum_{i=1}^{l}(\bar{x}_i - \bar{x})^2 n_i}{n} \qquad \textbf{(8.15)}$$

— межгрупповая дисперсия

Формула **(8.12)**, известная в статистике как «***правило сложения дисперсий***», имеет важное значение в статистическом анализе.

▷ **Пример 8.6.** Вычислить дисперсию, среднее квадратическое отклонение и коэффициент вариации распределения рабочих по выработке по данным табл. 8.1.

Р е ш е н и е. В примере **8.3** было получено $\bar{x} = 119{,}2(\%)$. По определению **(8.5)** дисперсия

$$s^2 = \frac{(97-119{,}2)^2\cdot 3 + (103-119{,}2)^2\cdot 7 + \ldots + (133-119{,}2)^2\cdot 10 + (139-119{,}2)^2\cdot 2}{100} =$$

$$= 87{,}48.$$

Среднее квадратическое отклонение $s = \sqrt{87{,}48} = 9{,}35(\%)$; коэффициент вариации по **(8.9)** $v = (9{,}35/119{,}2)100 = 7{,}8(\%)$.

Следует отметить, что вычисление дисперсии (особенно в случае, когда отклонения от средней $(x_i - \bar{x})^2$ выражаются нецелыми числами) бывает удобнее проводить по формуле **(8.10)**. Например, в данном примере вначале по **(8.11)** найдем

$$\overline{x^2} = \frac{97^2\cdot 3 + 103^2\cdot 7 + \ldots + 133^2\cdot 10 + 139^2\cdot 2}{100} = 14296{,}12.$$

Теперь по **(8.10)**

$$s^2 = \overline{x^2} - \bar{x}^2 = 14296{,}12 - 119{,}2^2 = 87{,}48. \blacktriangleright$$

▷ **Пример 8.7.** Имеются следующие данные о средних и дисперсиях заработной платы двух групп рабочих (табл. 8.3):

Таблица 8.3

Группа рабочих	*Число рабочих*	*Средняя заработная плата одного рабочего в группе (руб.)*	*Дисперсия заработной платы*
Работающие на одном станке	40	2400	180 000
Работающие на двух станках	60	3200	200 000

Найти общую дисперсию распределения рабочих по заработной плате и его коэффициент вариации.

Р е ш е н и е. Найдем общую среднюю по формуле **(8.5)**:

$$\bar{x} = \frac{2400 \cdot 40 + 3200 \cdot 60}{100} = 2880 \text{ (руб.)}.$$

Найдем среднюю групповых дисперсий по формуле **(8.13)**:

$$\overline{s_i^2} = \frac{180\,000 \cdot 40 + 200\,000 \cdot 60}{100} = 192\,000.$$

Найдем межгрупповую дисперсию по формуле **(8.15)**:

$$\delta^2 = \frac{(2400-2880)^2 \cdot 40 + (3200-2880)^2 \cdot 60}{100} = 153\,600.$$

Используя правило сложения дисперсий **(8.12)**, найдем общую дисперсию заработной платы и ее среднее квадратическое отклонение:

$$s^2 = 192\,000 + 153\,600 = 345\,600;\ s = \sqrt{345\,600} = 588 \text{ (руб.)}.$$

По формуле **(8.9)** коэффициент вариации

$$\tilde{v} = \frac{588}{2880} \cdot 100 = 20{,}4(\%). \blacktriangleright$$

8.4. Упрощенный способ расчета средней арифметической и дисперсии

Вычисление средней арифметической $\bar{x}$ и дисперсии s^2 вариационного ряда можно упростить, если использовать не первоначальные варианты x_i ($i = 1, 2, \ldots, m$), а новые варианты

$$u_i = \frac{x_i - c}{k}, \qquad \textbf{(8.16)}$$

где c и k — специально подобранные постоянные

Согласно свойствам **2** и **3** средней арифметической и дисперсии

$$\bar{u} = \left(\frac{\overline{x-c}}{k}\right) = \frac{\bar{x}-c}{k}, \qquad \textbf{(8.17)}$$

$$s_u^2 = s^2_{\frac{x-c}{k}} = \frac{s^2_{x-c}}{k^2} = \frac{s_x^2}{k^2},$$

откуда

$$\bar{x} = \bar{u}k + c \qquad \textbf{(8.18)}$$

и

$$s_x^2 = k^2 s_u^2. \qquad \textbf{(8.19)}$$

Учитывая **(8.10)**, а затем **(8.17)**, получим

$$s_x^2 = k^2\left(\overline{u^2} - \bar{u}^2\right) = k^2\overline{u^2} - k^2\bar{u}^2 = k^2\overline{u^2} - k^2\left(\frac{\bar{x}-c}{k}\right)^2 = k^2\overline{u^2} - (\bar{x}-c)^2.$$

Теперь, заменяя в **(8.18)** и **(8.19)** $\bar{u}$ и $\overline{u^2}$ их выражениями **(8.2)** и **(8.11)** через варианты u_i, получим

$$\bar{x} = \frac{\sum_{i=1}^{m} u_i n_i}{n} \cdot k + c, \qquad \textbf{(8.20)}$$

$$s_x^2 = \frac{\sum_{i=1}^{m} u_i^2 n_i}{n} \cdot k^2 - (\bar{x}-c)^2, \qquad \textbf{(8.21)}$$

где u_i определяются по **(8.16)**.

Формулы **(8.20)** и **(8.21)** дадут заметное упрощение расчетов, если в качестве постоянной k взять величину (ширину) интервала по x, а в качестве c — середину серединного интервала. Если серединных интервалов два (при четном числе интервалов), то в качестве c рекомендуется взять середину одного из этих интервалов, например, имеющего бóльшую частоту.

З а м е ч а н и е. Формулы **(8.20)** и **(8.21)** для $\bar{x}$ и s^2 носят технический, вспомогательный характер и позволяют рассчитать характеристики ряда по новым, условным вариантам. Основными же формулами, вытекающими из определения средней арифметической и дисперсии вариационного ряда и отражающими их сущность, остаются соответственно формулы **(8.3)** и **(8.7)**.

▷ **Пример 8.8.** Вычислить упрощенным способом среднюю арифметическую и дисперсию распределения рабочих по выработке по данным табл. 8.1.

Р е ш е н и е . Возьмем постоянную k, равную величине интервала, т.е. $k = 6$, и постоянную c, равную середине пятого (одного из двух серединных) интервала, т.е. $c = 121$. По **(8.16)** новые варианты $u_i = (x_i - 121)/6$.

Благодаря такому переходу получим вместо вариантов $x_i = 97, 103, 109, 115, 121, 127, 133$ «простые» варианты $u_i = -4, -3, -2, -1, 0, 1, 2, 3$.

Теперь для расчета x и s_x^2 по **(8.20)** и **(8.21)** необходимо найти суммы $\sum_{i=1}^{m} u_i n_i$ и $\sum_{i=1}^{m} u_i^2 n_i$. Их вычисление представим в табл. 8.4.

Таблица 8.4

i	*Интервалы* x	*Середина интервала* x_i	$u_i = \dfrac{x_i - 119}{10}$	n_i	$u_i n_i$	$u_i^2 n_i$	u_i+1	$(u_i+1)^2 n_i$
1	94,0–100,0	97	−4	3	−12	48	−3	27
2	100,0–106,0	103	−3	7	−21	63	−2	28
3	106,0–112,0	109	−2	11	−22	44	−1	11
4	112,0–118,0	115	−1	20	−20	20	0	0
5	118,0–124,0	121	0	28	0	0	1	28
6	124,0–130,0	127	1	19	19	19	2	76
7	130,0–136,0	133	2	10	20	40	3	90
8	136,0–142,0	139	3	2	6	18	4	32
	Σ	—	—	100	−30	252	—	292

В итоговой строке табл. 8.4 находим $\sum\limits_{i=1}^{8} u_i n_i = -30$, $\sum\limits_{i=1}^{8} u_i^2 n_i = 252$.

Последний столбец — контрольный. Если таблица составлена верно, то

$$\sum_{i=1}^{m}(u_i+1)^2 n_i = \sum_{i=1}^{m} u_i^2 n_i + 2\sum_{i=1}^{m} u_i n_i + n \quad (\text{где } n = \sum_{i=1}^{m} n_i).$$

В данном случае $\sum\limits_{i=1}^{8}(u_i+1)n_i = 292 = 252 + 2(-30) + 100$, т.е. расчеты проведены верно.

Теперь по **(8.20)** $\bar{x} = \dfrac{-30}{100}\cdot 6 + 121 = 119{,}2(\%)$, по **(8.21)** $s^2 = \dfrac{252}{100}\cdot 6^2 - (119{,}2-121)^2 = 87{,}48$. ▶

8.5. Начальные и центральные моменты вариационного ряда

Средняя арифметическая и дисперсия вариационного ряда являются частными случаями более общего понятия — моментов вариационного ряда.

Начальный момент $\tilde{\nu}_k$ *k-го порядка вариационного ряда*[1] определяется по формуле:

$$\tilde{\nu}_k = \frac{\sum\limits_{i=1}^{m} x_i^k n_i}{n}. \qquad (8.22)$$

[1] См. сноску на с. 292.

Очевидно, что $\tilde{\nu}_1 = \bar{x}$, т.е. средняя арифметическая является начальным моментом первого порядка вариационного ряда.

Центральный момент $\tilde{\mu}_k$ *k-го порядка вариационного ряда* определяется по формуле:

$$\tilde{\mu}_k = \frac{\sum\limits_{i=1}^{m}(x_i - \bar{x})^k n_i}{n}. \qquad (8.23)$$

С помощью моментов распределения можно описать не только среднюю тенденцию, рассеяние, но и другие особенности вариации признака.

Очевидно, в силу **(8.4)**, что $\tilde{\mu}_1 = 0$, а $\tilde{\mu}_2 = s^2$, т.е. центральный момент первого порядка для любого распределения равен нулю, а второго порядка является дисперсией вариационного ряда.

Коэффициентом асимметрии вариационного ряда называется число

$$\tilde{A} = \frac{\tilde{\mu}_3}{s^3} = \frac{\sum\limits_{i=1}^{m}(x_i - \bar{x})^3 n_i}{n s^3}. \qquad (8.24)$$

Если $\tilde{A} = 0$, то распределение имеет симметричную форму, т.е. варианты, равноудаленные от $\bar{x}$, имеют одинаковую частоту. При $\tilde{A} > 0$ $(\tilde{A} < 0)$ говорят о положительной (правосторонней) или отрицательной (левосторонней) асимметрии.

Эксцессом (или *коэффициентом эксцесса*) *вариационного ряда* называется число

$$\tilde{E} = \frac{\tilde{\mu}_4}{s^4} - 3 = \frac{\sum\limits_{i=1}^{m}(x_i - \bar{x})^4 n_i}{n s^4} - 3. \qquad (8.25)$$

Эксцесс является показателем «крутости» вариационного ряда по сравнению с нормальным распределением. Как отмечено выше (§ 4.7), эксцесс нормально распределенной случайной величины равен нулю.

Если $\tilde{E} > 0 (\tilde{E} < 0)$, то полигон вариационного ряда имеет более крутую (пологую) вершину по сравнению с нормальной кривой.

▷ **Пример 8.9.** Вычислить коэффициент асимметрии и эксцесс распределения рабочих по выработке по данным табл. 8.1.

Р е ш е н и е. Коэффициент асимметрии и эксцесс вариационного ряда, приведенного в табл. 8.1, найдем по формулам **(8.24)** и **(8.25)**:

$$\tilde{A} = \frac{(97-119{,}2)^3 \cdot 3 + (103-119{,}2)^3 \cdot 7 + \ldots + (139-119{,}2)^3 \cdot 2}{100 \cdot 9{,}35^3} = -0{,}302;$$

$$\tilde{E} = \frac{(97-119{,}2)^4 \cdot 3 + (103-119{,}2)^4 \cdot 7 + \ldots + (139-119{,}2)^4 \cdot 2}{100 \cdot 9{,}35^4} - 3 = -0{,}286.$$

В силу того, что коэффициент асимметрии $\tilde{A}$ отрицателен и близок нулю, распределение рабочих по выработке обладает незначительной левосторонней асимметрией, а поскольку эксцесс $\tilde{E}$ близок нулю, рассматриваемое распределение по крутости приближается к нормальной кривой. ▶

Средняя арифметическая $\bar{x}$, дисперсия s^2 и другие характеристики вариационного ряда являются статистическими аналогами математического ожидания $M(X)$, дисперсии σ^2 и соответствующих характеристик случайной величины X.

В табл. 8.5 приведено соответствие терминов (обозначений, формул) вариационного ряда и случайной величины. Подчеркнем, что вариационный ряд рассматривается в дальнейшем как одна из *реализаций* распределения признака (случайной величины)[1] X.

Таблица 8.5

Вариационный ряд		*Случайная величина*	
Обозначения, формулы	*Термин*	*Обозначения, формулы*	*Термин*
1	2	3	4
—	Дискретный ряд	—	Дискретная случайная величина
—	Интервальный ряд	—	Непрерывная случайная величина
x_i	Вариант	x_i, x	Значение случайной величины
w_i, w	Частость	p_i, p, P	Вероятность
—	Полигон, гистограмма	—	Полигон (многоугольник) распределения вероятностей, кривая распределения
$F_n(x)=w(X<x)$	Эмпирическая функция распределения	$F(x)=P(X<x)$	Функция распределения
$\bar{x} = \sum_{i=1}^{m} x_i w_i$	Средняя арифметическая	$a = M(X) = \sum_{i=1}^{n} x_i p_i$	Математическое ожидание*
$s^2 = \overline{(x-\bar{x})^2} = \sum_{i=1}^{m} (x_i - \bar{x})^2 w_i$	Дисперсия	$\sigma^2 = M[X - M(X)]^2 = \sum_{i=1}^{n} (x_i - a)^2 p_i$	Дисперсия*
$s = \sqrt{s^2}$	Среднее квадратическое отклонение	$\sigma = \sqrt{D(X)} = \sqrt{\sigma^2}$	Среднее квадратическое отклонение
$\tilde{Mo}$	Мода	$Mo(X)$	Мода
$\tilde{Me}$	Медиана	$Me(X)$	Медиана
$\tilde{\nu}_k = \sum_{i=1}^{m} x_i^k w_i$	Начальный момент k-го порядка	$\nu_k = \sum_{i=1}^{n} x_i^k p_i$	Начальный момент k-го порядка*
$\tilde{\mu}_k = \sum_{i=1}^{m} (x_i - \bar{x})^k w_i$	Центральный момент k-го порядка	$\mu_k = \sum_{i=1}^{n} [x_i - M(X)]^k p_i$	Центральный момент k-го порядка*
$\tilde{A} = \tilde{\mu}_3 / s^3$	Коэффициент асимметрии	$A = \mu_3 / \sigma^3$	Коэффициент асимметрии
$\tilde{E} = \tilde{\mu}_4 / s^4 - 3$	Эксцесс	$E = \mu_4 / \sigma^4 - 3$	Эксцесс

* Формула приведена для дискретной случайной величины.

[1] Если для характеристик вариационного ряда используются те же буквенные выражения, что и для случайной величины, то обозначения этих характеристик дополняются знаком ~ («тильда»).

Окончание табл. 8.5

Упражнения

В **примерах 8.10—8.12** дано распределение признака X (случайной величины X), полученной по n наблюдениям. Необходимо[1]: 1) построить полигон (гистограмму), кумуляту и эмпирическую функцию распределения X; 2) найти: а) среднюю ариф-

[1] При наличии открытых интервалов значений X типа «менее x_1» или «свыше x_n» для проведения расчетов их условно заменяют интервалами той же ширины k, т.е. $(x_1 - k, x_1)$ или $(x_n, x_n + k)$.

метическую $\bar{x}$; б) медиану *Me* и моду *Mo*; в) дисперсию s^2, среднее квадратическое отклонение s и коэффициент вариации $\tilde{v}$; г) начальные $\tilde{\nu}_k$ и центральные $\tilde{\mu}_k$ моменты k-го порядка ($k = 1, 2, 3, 4$); д) коэффициент асимметрии $\tilde{A}$ и эксцесс $\tilde{E}$.

8.10. X — число сделок на фондовой бирже за квартал; $n = 400$ (инвесторов).

x_i	0	1	2	3	4	5	6	7	8	9	10
n_i	146	97	73	34	23	10	6	3	4	2	2

8.11. X — месячный доход жителя региона (в руб.); $n = 1000$ (жителей).

x_i	Менее 500	500—1000	1000—1500	1500—2000	2000—2500	Свыше 2500
n_i	58	96	239	328	147	132

8.12. X — удой коров на молочной ферме за лактационный период (в ц); $n = 100$ (коров).

x_i	4—6	6—8	8—10	10—12	12—14	14—16	16—18	18—20	20—22	22—24	24—26
n_i	1	3	6	11	15	20	14	12	10	6	2

8.13. В таблице приведено распределение 50 рабочих по производительности труда X (единиц за смену), разделенных на две группы: 30 и 20 человек.

	Прошедшие техническое обучение (группа 1)					*Не прошедшие техническое обучение (группа 2)*				
x_i	85	34	96	102	103	63	69	83	89	106
n_i	2	5	11	8	4	2	6	8	3	1

Вычислить общие и групповые средние и дисперсии и убедиться в справедливости правила сложения дисперсий.

Глава 9 Основы математической теории выборочного метода

9.1. Общие сведения о выборочном методе

В практике статистических наблюдений различают два вида наблюдений: *сплошное*, когда изучаются все объекты (элементы, единицы) совокупности, и несплошное, *выборочное*, когда изучается часть объектов. Примером сплошного наблюдения является перепись населения, охватывающая все население страны. Выборочными наблюдениями является, например, проводимые социологические исследования, охватывающие часть населения страны, области, района и т.д.

Вся подлежащая изучению совокупность объектов (наблюдений) называется ***генеральной совокупностью.*** В математической статистике *понятие генеральной совокупности трактуется как совокупность всех мыслимых наблюдений, которые могли бы быть произведены при данном реальном комплексе условий*, и в этом смысле его не следует смешивать с реальными совокупностями, подлежащими статистическому изучению. Так, обследовав даже все предприятия подотрасли по определенным технико-экономическим показателям, мы можем рассматривать обследованную совокупность лишь как представителя гипотетически возможной более широкой совокупности предприятий, которые могли бы функционировать в рамках того же реального комплекса условий.

Понятие генеральной совокупности в определенном смысле аналогично понятию случайной величины (закону распределения вероятностей, вероятностному пространству), так как полностью обусловлено определенным комплексом условий.

Та часть объектов, которая отобрана для непосредственного изучения из генеральной совокупности, называется ***выборочной совокупностью, или выборкой.*** Числа объектов (наблюдений) в генеральной или выборочной совокупности называются их *объемами.* Генеральная совокупность может иметь как конечный, так и бесконечный объем.

Выборку можно рассматривать как некий эмпирический аналог генеральной совокупности. *Сущность выборочного метода состоит в том, чтобы по некоторой части генеральной совокупности (по выборке) выносить суждение о ее свойствах в целом.*

Отметим **преимущества выборочного метода** наблюдения по сравнению со сплошным:

- позволяет существенно экономить *затраты ресурсов* (материальных, трудовых, временных);
- является *единственно возможным* в случае бесконечной генеральной совокупности или в случае, когда исследование связано с уничтожением наблюдаемых объектов (например, исследование долговечности электрических лампочек, предельных режимов работы приборов и т.п.);
- при тех же затратах ресурсов дает возможность проведения *углубленного исследования* за счет расширения программы исследования;
- позволяет снизить *ошибки регистрации*, т.е. расхождения между истинным и зарегистрированным значениями признака.

Основной недостаток выборочного метода — ошибки исследования, называемые *ошибками репрезентативности* (*представительства*), о которых речь пойдет ниже.

Однако неизбежные ошибки, возникающие при выборочном методе исследования в связи с изучением только части объектов, могут быть заранее оценены и посредством правильной организации выборки сведены к практически незначимым величинам. Между тем использование сплошного наблюдения даже там, где это принципиально возможно, не говоря уже о росте трудоемкости, стоимости и увеличении необходимого времени, часто приводит к тому, что каждое отдельное наблюдение поневоле проводится с меньшей точностью. А это уже сопряжено с неустранимыми ошибками и в конечном счете может привести к снижению точности сплошного наблюдения по сравнению с выборочным.

Чтобы по данным выборки иметь возможность судить о генеральной совокупности, она должна быть отобрана ***случайно.*** Случайность отбора элементов в выборку достигается соблюдением принципа равной возможности всем элементам генеральной совокупности быть отобранными в выборку. На практике это достигается тем, что извлечение элементов в выборку проводится путем жеребьевки (лотереи) или с помощью случайных чисел, имеющихся в специальных таблицах или вырабатываемых ЭВМ с помощью датчика случайных чисел.

Выборка называется *репрезентативной* (*представительной*), если она достаточно хорошо воспроизводит генеральную совокупность.

Различают следующие виды выборок:

- *собственно-случайная выборка*, образованная случайным выбором элементов без расчленения на части или группы;
- *механическая выборка*, в которую элементы из генеральной совокупности отбираются через определенный интервал. Например, если объем выборки должен составлять 10% (10%-ная выборка), то отбирается каждый 10-й ее элемент и т.д.;
- *типическая* (*стратифицированная*) *выборка*, в которую случайным образом отбираются элементы из типических групп, на которые по некоторому признаку разбивается генеральная совокупность;
- *серийная* (*гнездовая*) *выборка*, в которую случайным образом отбираются не элементы, а целые группы совокупности (серии), а сами серии подвергаются сплошному наблюдению.

Используют два способа образования выборки:

- *повторный отбор* (по схеме возвращенного шара), когда каждый элемент, случайно отобранный и обследованный, возвращается в общую совокупность и может быть повторно отобран;
- *бесповторный отбор* (по схеме невозвращенного шара), когда отобранный элемент не возвращается в общую совокупность.

Математическая теория выборочного метода основывается на анализе собственно-случайной выборки. Рассмотрением этой выборки мы и ограничимся.

Обозначим:

x_i	—	значения признака (случайной величины X);
N и n	—	объемы генеральной и выборочной совокупностей;
N_i и n_i	—	число элементов генеральной и выборочной совокупностей со значением признака x_i;
M и m	—	число элементов генеральной и выборочной совокупностей, обладающих данным признаком.

Средние арифметические распределения признака в генеральной и выборочной совокупностях называются соответственно *генеральной* и *выборочной средними*, а дисперсии этих распределений — *генеральной* и *выборочной дисперсиями*. Отношение числа элементов генеральной и выборочной совокупностей, обладающих некоторым признаком A, к их объемам, называются

соответственно *генеральной* и *выборочной долями*. Все формулы сведем в таблицу

Таблица 9.1

Наименование характеристики	*Генеральная совокупность*		*Выборка*	
Средняя	$\bar{x}_0 = \frac{\sum_{i=1}^{m} x_i N_i}{N}$	**(9.1)**	$\bar{x} = \frac{\sum_{i=1}^{m} x_i n_i}{n}$	**(9.2)**
Дисперсия	$\sigma^2 = \frac{\sum_{i=1}^{m} (x_i - \bar{x}_0)^2 N_i}{N}$	**(9.3)**	$s^2 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 n_i}{n}$	**(9.4)**
Доля	$p = \frac{M}{N}$	**(9.5)**	$w = \frac{m}{n}$	**(9.6)**

Замечание. В случае бесконечной генеральной совокупности $(N = \infty)$ под генеральными средней и дисперсией понимается соответственно математическое ожидание $a = \bar{x}_0$ и дисперсия σ^2 распределения признака X (генеральной совокупности), а под генеральной долей p — вероятность данного события.

Важнейшей ***задачей выборочного метода*** *является оценка параметров (характеристик) генеральной совокупности по данным выборки.*

Теоретическую основу применимости выборочного метода составляет *закон больших чисел*, согласно которому при неограниченном увеличении объема выборки практически достоверно, что случайные выборочные характеристики как угодно близко приближаются (сходятся по вероятности) к определенным параметрам генеральной совокупности.

9.2. Понятие оценки параметров

Сформулируем задачу оценки параметров в общем виде. Пусть распределение признака X — генеральной совокупности — задается функцией вероятностей $\varphi(x_i,\theta) = P(X = x_i)$ (для дискретной случайной величины X) или плотностью вероятности $\varphi(x,\theta)$ (для непрерывной случайной величины X), которая содержит

неизвестный параметр θ. Например, это параметр λ в распределении Пуассона или параметры a и σ^2 для нормального закона распределения и т.д.

Для вычисления параметра θ исследовать все элементы генеральной совокупности не представляется возможным. Поэтому о параметре θ пытаются судить по выборке, состоящей из значений (вариантов) $x_1, x_2,..., x_n$. Эти значения можно рассматривать как частные значения (реализации) n независимых случайных величин $X_1, X_2,..., X_n$, каждая из которых имеет тот же закон распределения, что и сама случайная величина X.

Определение. ***Оценкой*** $\tilde{\theta}_n$ *параметра* θ *называют всякую функцию результатов наблюдений над случайной величиной* X *(иначе — статистику), с помощью которой судят о значении параметра* θ:

$$\tilde{\theta}_n = \tilde{\theta}_n(X_1, X_2,..., X_n).$$

Поскольку $X_1, X_2,..., X_n$ — случайные величины, то и *оценка* $\tilde{\theta}_n$ (в отличие от оцениваемого параметра θ — величины неслучайной, детерминированной) *является случайной величиной, зависящей от закона распределения случайной величины* X *и числа* n.

Всегда существует множество функций от результатов наблюдений $X_1, X_2,..., X_n$ (от n «экземпляров» случайной величины X), которые можно предложить в качестве оценки параметра θ. Например, если параметр θ является математическим ожиданием случайной величины X, т.е. генеральной средней x_0, то в качестве его оценки $\tilde{\theta}_n$ по выборке можно взять: среднюю арифметическую результатов наблюдений — выборочную среднюю $\bar{x}$, моду $\tilde{Mo}$, медиану $\tilde{Me}$, полусумму наименьшего и наибольшего значений по выборке, т.е. $(x_{\min}+x_{\max})/2$ и т.д. Какими свойствами должна обладать оценка $\tilde{\theta}_n$, чтобы в каком-то смысле быть «доброкачественной» оценкой?

Назвать «наилучшей» оценкой такую, которая наиболее близка к истинному значению оцениваемого параметра, невозможно, так как выше отмечено, что $\tilde{\theta}_n$ — случайная величина, поэтому невозможно предсказать индивидуальное значение оценки в данном частном случае. Так что *о качестве оценки следует судить не по индивидуальным ее значениям, а лишь по распре-*

делению ее значений в большой сети испытаний, т.е. *по выборочному распределению оценки.* Если значения оценки $\tilde{\theta}_n$ концентрируются около истинного значения параметра θ, т.е. основная часть массы выборочного распределения оценки сосредоточена в малой окрестности оцениваемого параметра θ, то с большой вероятностью можно считать, что оценка $\tilde{\theta}_n$ отличается от параметра θ лишь на малую величину. Поэтому, чтобы значение $\tilde{\theta}_n$ было близко к θ, надо, очевидно, потребовать, чтобы *рассеяние случайной величины* $\tilde{\theta}_n$ относительно θ, выражаемое, например, математическим ожиданием квадрата отклонения оценки от оцениваемого параметра $M\left(\tilde{\theta}_n-\theta\right)^2$, *было по возможности меньшим.* Таково основное условие, которому должна удовлетворять «наилучшая» оценка.

Рассмотрим наиболее важные **свойства оценок**.

Определение. *Оценка* $\tilde{\theta}_n$ *параметра* θ *называется* ***несмещенной****, если ее математическое ожидание равно оцениваемому параметру*, т.е.

$$M\left(\tilde{\theta}_n\right)=\theta.$$

В противном случае оценка называется *смещенной.*

Если это равенство не выполняется, то оценка $\tilde{\theta}_n$, полученная по разным выборкам, будет в среднем либо завышать значение θ (если $M\left(\tilde{\theta}_n\right)>\theta$), либо занижать его (если $M\left(\tilde{\theta}_n\right)<\theta$). Таким образом, *требование несмещенности гарантирует отсутствие систематических ошибок при оценивании.*

Замечание. На первый взгляд, приведенное выше определение любой оценки, как всякой функции результатов наблюдений, было бы более естественным и не таким расплывчатым, если бы в нем содержалось условие $M\left(\tilde{\theta}_n\right)=\theta$. К сожалению, этого сделать нельзя, так как практически важные оценки оказываются смещенными, хотя и слабо.

Если при конечном объеме выборки n $M\left(\tilde{\theta}_n\right)\neq\theta$, т.е. *смещение* оценки $b(\tilde{\theta}_n)=M(\tilde{\theta}_n)-\theta\neq 0$, но $\lim\limits_{n\to\infty} b(\theta_n)=0$, то такая оценка $\tilde{\theta}_n$ называется ***асимптотически несмещенной.***

Определение. *Оценка* $\tilde{\theta}_n$ *параметра* θ *называется* ***состоятельной****, если она удовлетворяет закону больших чисел, т.е. сходится по вероятности к оцениваемому параметру*:

$$\lim_{n\to\infty} P\left(\left|\tilde{\theta}_n-\theta\right|<\varepsilon\right)=1, \qquad (9.7)$$

или

$$\tilde{\theta}_n \xrightarrow[n\to\infty]{\mathcal{P}} \theta.$$

В случае использования состоятельных оценок оправдывается увеличение объема выборки, так как при этом становятся маловероятными значительные ошибки при оценивании. Поэтому *практический смысл имеют только состоятельные оценки.* Если оценка состоятельна, то практически достоверно, что при достаточно большом n $\tilde{\theta}_n\approx\theta$.

Если оценка $\tilde{\theta}_n$ *параметра* θ *является несмещенной, а ее дисперсия* $\sigma^2_{\tilde{\theta}_n}\to 0$ *при* $n\to\infty$, *то оценка* $\tilde{\theta}_n$ *является и состоятельной.* Это непосредственно вытекает из неравенства Чебышева:

$$P\left(\left|\tilde{\theta}_n-\theta\right|<\varepsilon\right)\geq 1-\frac{\sigma^2_{\tilde{\theta}_n}}{\varepsilon^2}.$$

Так, например, выборочная средняя $\bar{x}$ является несмещенной и состоятельной оценкой генеральной средней $\overline{x_0}$ (дисперсия $\sigma^2_{\bar{x}}\to 0$ при $n\to\infty$, см. § 9.4), а отдельное выборочное наблюдение $X_k\left(k=1,2,...,n\right)$ — несмещенной $\left(M\left(X_k\right)=M\left(X\right)=\bar{x}_0\right)$, но не состоятельной оценкой генеральной средней, так как ее дисперсия $\sigma^2\left(X_i\right)=\sigma^2\left(X\right)=\sigma^2$ постоянна и не уменьшается с ростом n.

Определение. *Несмещенная оценка* $\tilde{\theta}_n$ *параметра* θ *называется* ***эффективной****, если она имеет наименьшую дисперсию среди всех возможных несмещенных оценок параметра* θ, *вычисленных по выборкам одного и того же объема n.*

Так как для несмещенной оценки[1] $M\left(\tilde{\theta}_n-\theta\right)^2$ есть ее дисперсия $\sigma^2_{\tilde{\theta}_n}$, то эффективность является решающим свойством, определяющим качество оценки.

[1] Для смещенной оценки, как нетрудно показать, $M\left(\tilde{\theta}_n-\theta\right)^2=\sigma^2_{\tilde{\theta}_n}+b^2(\tilde{\theta}_n)$, где $b(\tilde{\theta}_n)$ — смещение оценки.

Эффективность оценки $\tilde{\theta}_n$ определяют отношением:

$$e = \frac{\sigma^2_{\tilde{\theta}_n^э}}{\sigma^2_{\tilde{\theta}_n}}, \qquad (9.8)$$

где $\sigma^2_{\tilde{\theta}_n^э}$ и $\sigma^2_{\tilde{\theta}_n}$ — соответственно дисперсии эффективной и данной оценок. Чем ближе e к 1, тем эффективнее оценка. *Если* $e \to 1$ *при* $n \to \infty$, *то такая оценка называется* ***асимптотически эффективной***.

На практике в целях упрощения расчетов используются оценки, не обладающие высокой эффективностью. Так, например, генеральную среднюю $\bar{x}_0$ часто оценивают медианой $\tilde{M}e$ выборки, в то время как эффективной оценкой $\bar{x}_0$ является выборочная средняя $\bar{x}$ (§ 9.5). При нормальном распределении признака в генеральной совокупности можно показать, что асимптотическая эффективность этой оценки, т.е. $e(\tilde{M}e) = 2/\pi = 0{,}64$ при $n \to \infty$. Это означает, что для получения той же точности и надежности оценки генеральной средней по выборочной средней нужно использовать лишь 64% объема выборки, взятого при оценке по медиане.

Другой пример. В практике статистического контроля качества продукции для оценки генерального среднего квадратического отклонения σ широко используют оценку $s_R = R/d_n$, где $R = x_{\max} - x_{\min}$ — вариационный размах, d_n — коэффициент, зависящий от объема выборки n. При малых n эффективность оценки s_R достаточно высока, но с увеличением n быстро падает. Поэтому удовлетворительная оценка σ с помощью s_R может быть достигнута лишь при $n < 10$.

В качестве статистических оценок параметров генеральной совокупности желательно использовать оценки, удовлетворяющие одновременно требованиям несмещенности, состоятельности и эффективности. Однако достичь этого удается не всегда. Может оказаться, что для простоты расчетов целесообразно использовать незначительно смещенные оценки или оценки, обладающие большей дисперсией по сравнению с эффективными оценками, и т.п.

9.3. Методы нахождения оценок

Рассмотрим основные методы нахождения оценок.

Согласно **методу моментов**, предложенному К. Пирсоном, *определенное количество выборочных моментов (начальных* $\tilde{\nu}_k$ *или центральных* $\tilde{\mu}_k$, *или тех и других) приравнивается к соответствующим теоретическим моментам распределения* (ν_k или μ_k) *случайной величины X.* Напомним, что выборочные моменты $\tilde{\nu}_k$ и $\tilde{\mu}_k$ определяются по формулам **(8.22)** и **(8.23)**, а соответствующие им теоретические моменты — по формулам **(3.32)**—**(3.35)**:

$$\nu_k = \sum_{i=1}^{n} x_i^k p_i, \quad \mu_k = \sum_{i=1}^{n} (x_i - a)^k p_i$$

(для дискретной случайной величины с функцией вероятностей $p_i = \varphi(x_i, \theta)$),

$$\nu_k = \int_{-\infty}^{+\infty} x^k \varphi(x, \theta) dx, \quad \mu_k = \int_{-\infty}^{+\infty} (x - a)^k \varphi(x, \theta) dx$$

(для непрерывной случайной величины с плотностью вероятностей $\varphi(x, \theta)$), где $a = M(X)$ — см. § 3.7.

▷ **Пример 9.1.** Найти оценку метода моментов для параметра λ закона Пуассона.

Р е ш е н и е. В данном случае для нахождения единственного параметра λ достаточно приравнять теоретический ν_1 и эмпирический $\tilde{\nu}_1$ начальные моменты первого порядка. ν_1 — математическое ожидание случайной величины X. В § 4.2 установлено, что для случайной величины, распределенной по закону Пуассона, $M(X) = \lambda$. Момент $\tilde{\nu}_1$ согласно **(8.22)** равен x. Следовательно, *оценка метода моментов параметра* λ *закона Пуассона есть выборочная средняя* $\bar{x}$. ▶

Оценки метода моментов обычно *состоятельны*, однако по эффективности они не являются «наилучшими», их эффективности $e(\tilde{\theta}_n)$ часто значительно меньше единицы. Тем не менее, метод моментов часто используется на практике, так как приводит к сравнительно простым вычислениям.

Основным методом получения оценок параметров генеральной совокупности по данным выборки является **метод максимального (наибольшего) правдоподобия**, предложенный Р. Фишером.

Основу метода составляет ***функция правдоподобия***, *выражающая плотность вероятности (вероятность) совместного появления результатов выборки* $x_1, x_2, \ldots, x_n$:

$$L(x_1, x_2, \ldots, x_i, \ldots, x_n; \theta) = \varphi(x_1, \theta) \cdot \varphi(x_2, \theta) \ldots \varphi(x_i, \theta) \ldots \varphi(x_n, \theta),$$

или
$$L(x_1, x_2, \ldots, x_i, \ldots, x_n; \theta) = \prod_{i=1}^{n} \varphi(x_i, \theta).$$

Согласно методу максимального правдоподобия в качестве оценки неизвестного параметра θ *принимается такое значение* $\tilde{\theta}_n$, *которое максимизирует функцию* L. Естественность подобного подхода к определению статистических оценок вытекает из смысла функции правдоподобия, которая при каждом фиксированном значении параметра θ является мерой правдоподобности получения наблюдений $x_1, x_2, \ldots, x_n$. И оценка $\tilde{\theta}_n$ такова, что имеющиеся у нас наблюдения $x_1, x_2, \ldots, x_n$ являются наиболее правдоподобными.

Нахождение оценки $\tilde{\theta}_n$ упрощается, если максимизировать не саму функцию L, а $\ln L$, поскольку максимум обеих функций достигается при одном и том же значении θ. Поэтому для отыскания оценки параметра θ (одного или нескольких) надо решить *уравнение (систему уравнений) правдоподобия*, получаемое приравниванием производной (частных производных) нулю по параметру (параметрам) θ:

$$\frac{d \ln L}{d\theta} = 0 \quad \text{или} \quad \frac{1}{L}\frac{dL}{d\theta} = 0, \qquad \textbf{(9.9)}$$

а затем отобрать то решение, которое обращает функцию $\ln L$ в максимум.

▷ **Пример 9.2.** Найти оценку метода максимального правдоподобия для вероятности p наступления некоторого события A по данному числу m появления этого события в n независимых испытаниях.

Р е ш е н и е. Составим функцию правдоподобия:

$$L = (x_1, x_2, \ldots, x_n; p) = \underbrace{pp \ldots p}_{m} \underbrace{(1-p)(1-p) \ldots (1-p)}_{n-m}, \text{ или}$$

$L = p^m (1-p)^{n-m}$. Тогда $\ln L = m \ln p + (n-m) \ln (1-p)$ и согласно уравнению **(9.9)**

$$\frac{d \ln L}{dp} = \frac{m}{p} - \frac{n-m}{1-p}, \text{ откуда } \tilde{p} = \frac{m}{n}$$ (можно показать, что при $\tilde{p} = m/n$ выполняется достаточное условие экстремума функции L).

Таким образом, *оценкой метода максимального правдоподобия вероятности* p *события* A *будет частость* $w = \frac{m}{n}$ *этого события.* ▶

▷ **Пример 9.3.** Найти оценки метода максимального правдоподобия для параметров a и σ^2 нормального закона распределения по данным выборки.

Р е ш е н и е. Плотность вероятности нормально распределенной случайной величины

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

Тогда функция правдоподобия имеет вид:

$$L(x_1, x_2, \ldots, x_n; a, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-a)^2}{2\sigma^2}} = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} e^{\frac{-\sum_{i=1}^{n}(x_i-a)^2}{2\sigma^2}}$$

Логарифмируя, получим:

$$\ln L = -\frac{n}{2}\left(\ln \sigma^2 + \ln(2\pi)\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - a)^2 .$$

Для нахождения параметров a и σ^2 надо приравнять нулю частные производные по параметрам a и σ^2, т.е. решить систему уравнений правдоподобия:

$$\begin{cases} \dfrac{\partial \ln L}{\partial a} = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i - a) = 0, \\ \dfrac{\partial \ln L}{\partial \sigma^2} = \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(x_i - a)^2 - \dfrac{n}{2\sigma^2} = 0, \end{cases}$$

откуда оценки максимального правдоподобия равны:

$$\tilde{a} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}, \qquad \tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = s^2.$$

Таким образом, *оценками метода максимального правдоподобия математического ожидания a и дисперсии σ^2 нормально распределенной случайной величины являются соответственно выборочная средняя $\bar{x}$ и выборочная дисперсия s^2.* ▶

Важность метода максимального правдоподобия связана с его оптимальными свойствами. Так, *если для параметра θ существует эффективная оценка $\tilde{\theta}_n^э$, то оценка максимального правдоподобия единственная и равна $\tilde{\theta}_n^э$.* Кроме того, при достаточно общих условиях *оценки максимального правдоподобия являются состоятельными, асимптотически несмещенными, асимптотически эффективными и имеют асимптотически нормальное распределение.*

Основной недостаток метода максимального правдоподобия — трудность вычисления оценок, связанных с решением уравнений правдоподобия, чаще всего нелинейных. Существенно также и то, что для построения оценок максимального правдоподобия и обеспечения их «хороших» свойств необходимо точное знание типа анализируемого закона распределения $\varphi(x,\theta)$, что во многих случаях оказывается практически нереальным.

Метод наименьших квадратов — один из наиболее простых приемов построения оценок. Суть его заключается в том, что *оценка определяется из условия минимизации суммы квадратов отклонений выборочных данных от определяемой оценки.*

▷ **Пример 9.4.** Найти оценку метода наименьших квадратов $\tilde{\theta}_n$ для генеральной средней $\theta = x_0$.

Р е ш е н и е. Согласно методу наименьших квадратов найдем оценку $\tilde{\theta}_n$ из условия минимизации суммы:

$$u = \sum_{i=1}^{n} (x_i - \theta)^2 \to \min .$$

Используя необходимое условие экстремума, приравняем нулю производную

$$\frac{du}{d\theta} = -2\sum_{i=1}^{n}(x_i - \theta) = 0, \quad \text{откуда} \quad \sum_{i=1}^{n} x_i - \theta n = 0$$

и $\tilde{\theta}_n = \dfrac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$, т.е. *оценка метода наименьших квадратов генеральной средней $\bar{x}_0$ есть выборочная средняя $\bar{x}$.* ▶

Заметим, что полученная в примере **9.4** оценка метода наименьших квадратов $\bar{x}$ для генеральной средней $\bar{x}_0$ совпала с оценкой метода максимального правдоподобия для математического ожидания $a = \bar{x}_0$ для нормально распределенной случайной величины (см. пример **9.3**).

Метод наименьших квадратов получил самое широкое распространение в практике статистических исследований, так как, во-первых, *не требует знания закона распределения выборочных данных*; во-вторых, достаточно хорошо разработан в плане вычислительной реализации.

Применение метода наименьших квадратов в задачах корреляционного и регрессионного анализа рассмотрено в гл. 12 и 13.

9.4. Оценка параметров генеральной совокупности по собственно-случайной выборке

Оценка генеральной доли. Пусть генеральная совокупность содержит N элементов, из которых M обладает некоторым признаком A. Следует найти «наилучшую» оценку генеральной доли $p = \dfrac{M}{N}$.

Рассмотрим в качестве такой возможной оценки параметра p его статистический аналог — выборочную долю $w = \dfrac{m}{n}$.

а) **Выборка повторная**

Выборочную долю можно представить как среднюю арифметическую n альтернативных случайных величин[1] $X_1, X_2, \ldots, X_k, \ldots, X_n$, т.е. $w = \dfrac{\sum_{k=1}^{n} X_k}{n}$, где каждая случайная величина X_k (k=1,2,...,n) выражает число появлений признака в k-м элементе выборки (т.е. при наличии признака X_k=1, при его отсутствии X_k=0) и имеет один и тот же закон распределения:

[1] В учебнике случайные величины, как правило, обозначаются прописными буквами, а их значения — строчными. Выборочные среднюю и долю везде обозначаем для простоты строчными буквами, соответственно $\bar{x}$ и w. При этом следует понимать, что до *проведения наблюдений*, когда заранее неизвестно, какими они будут, $\bar{x}$ и w *рассматриваем как* с л у ч а й н ы е *величины*; *после проведения наблюдений*, когда получены их конкретные значения, — как н е с л у ч а й н ы е *величины*.

x_i	0	1
p_i	$\frac{N-M}{N}$	$\frac{M}{N}$

(9.10)

Действительно, вероятность того, что 1-й отобранный в выборку элемент обладает признаком A, согласно классическому определению вероятности равна $p(X_1=1)=\frac{M}{N}$, так как из общего числа N элементов генеральной совокупности M элементов обладают признаком A. Аналогично вероятность того, что 1-й элемент не обладает признаком A, равна $p(X_1=0)=\frac{N-M}{N}$. Так как выборка повторная, и каждый отобранный и обследованный элемент вновь возвращается в исходную совокупность, восстанавливая всякий раз ее первоначальные состав и объем, то вероятности $p(X_k=0)$ и $p(X_k=1)$ остаются теми же для любого элемента выборки, и закон распределения X_k ($k=1,2,...,n$) один и тот же — **(9.10)**.

Случайные величины X_1, $X_2,...,X_k,...,X_n$ *независимы*, так как независимы любые события $X_k=0$, $X_k=1$ ($k=1,2,...,n$) и их комбинации. Например, независимы события $X_1=1$ и $X_2=1$, так как $p_{X_1=1}(X_2=1) = p(X_2=1)=\frac{M}{N}$, т.е. вероятность того, что 2-й отобранный в выборку элемент обладает признаком A, не меняется в зависимости от того, обладал признаком A 1-й элемент или нет, и т.д.

Теорема. *Выборочная доля* $w=\frac{m}{n}$ *повторной выборки есть несмещенная и состоятельная оценка генеральной доли* $p=\frac{M}{N}$, *причем ее дисперсия*

$$\sigma_w^2=\frac{pq}{n}, \qquad \textbf{(9.11)}$$

где $q = 1 - p$.

□ Докажем вначале несмещенность оценки w. Математическое ожидание и дисперсия частости события в n

независимых испытаниях, в каждом из которых оно может наступить с одной и той же вероятностью p, равны соответственно

$$M(w)=p,\quad D(w)=\sigma_w^2=\frac{pq}{n},$$

где $q = 1-p$ (см. § 4.1).

Так как вероятность того, что любой отобранный в выборку элемент обладает признаком A, есть генеральная доля p, то из первого равенства вытекает, что частость или выборочная доля w есть несмещенная оценка генеральной доли p.

Осталось доказать состоятельность оценки $w=\frac{m}{n}$, которая следует непосредственно из теоремы Бернулли (§ 6.4):

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n}-p\right|\le\varepsilon\right)=1,$$

или $$w\xrightarrow[n\to\infty]{\mathcal{P}} p.$$ ■

б) **Выборка бесповторная**

В случае бесповторной выборки случайные величины X_1, $X_2,...,X_n$ будут *зависимыми*. Рассмотрим, например, события $X_1=1$ и $X_2=1$. Теперь вероятность $p_{X_1=1}(X_2=1)=\frac{M-1}{N-1}$, так как отобранный элемент (в случае бесповторной выборки) в исходную совокупность не возвращается, то в ней остается всего $N-1$ элементов, из которых обладающих признаком A $M-1$. Эта вероятность $p_{X_1=1}(X_2=1)$ не равна $p(X_2=1)=\frac{M}{N}$, т.е. события $X_1=1$ и $X_2=1$ — зависимые. Аналогично будут зависимы любые события $X_k=1$, $X_k=0$ ($k=1,2,...,n$), а значит, зависимы случайные величины X_1, $X_2,...,X_k,...,X_n$.

Однако и для бесповторной выборки выборочная доля является «хорошей» оценкой. Об этом свидетельствует следующая теорема.

Теорема. *Выборочная доля* $w=\frac{m}{n}$ *бесповторной выборки есть несмещенная и состоятельная оценка генеральной доли* $p=\frac{M}{N}$, *причем ее дисперсия*

$$\sigma'^2_w = \frac{pq}{n}\left(\frac{N-n}{N-1}\right) \approx \frac{pq}{n}\left(1-\frac{n}{N}\right), \qquad \textbf{(9.12)}$$

где $q = 1 - p$.

□ Очевидно, что и для бесповторной выборки $M(w) = p$, т.е. w — несмещенная оценка для генеральной доли $p=M/N$. Это связано с тем, что математическое ожидание суммы любых случайных величин равно сумме их математических ожиданий (в том числе суммы зависимых случайных величин, каковой является выборочная доля w бесповторной выборки).

Найдем дисперсию выборочной доли для бесповторной выборки:

$$\sigma'^2_w = \sigma'^2\left(\frac{m}{n}\right) = \frac{1}{n^2}\sigma'^2(m) = \frac{1}{n^2}\left[n\frac{M}{N-1}\left(1-\frac{M}{N}\right)\left(1-\frac{n}{N}\right)\right] =$$

$$= \frac{1}{n}\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1} = \frac{pq}{n}\frac{N-n}{N-1},$$

где $p = M/N$, $q = 1 - M/N$, т.е. верна формула **(9.12)** (при выводе формулы для σ'^2_w использовали то, что случайная величина $X=m$ в случае бесповторной выборки имеет гипергеометрическое распределение (см. § 4.4), и ее дисперсия определяется по формуле **(4.16)**). ■

Для того чтобы легче было понять формулу **(9.12)**, рассмотрим ее частные случаи и убедимся в справедливости этой формулы:

1. При $n \ll N$ $\sigma'^2_w = \frac{pq}{n}\left(\frac{N-n}{N-1}\right) \approx \frac{pq}{n} = \sigma^2_w$, т.е. если объем выборки значительно меньше объема генеральной совокупности, то выборка практически не отличается от повторной и, естественно, что дисперсии выборочной доли σ^2_w и σ'^2_w приближенно равны.

2. При $n=N$ $\sigma'^2_w = 0$, т.е. если предположить, что объем выборки равен объему генеральной совокупности, то выборочная доля будет равна генеральной доле и ее дисперсия будет равна нулю.

▷ **Пример 9.5.** Найти несмещенную и состоятельную оценку доли рабочих цеха с выработкой не менее 124% по выборке, представленной в табл. 8.1.

Р е ш е н и е. Несмещенной и состоятельной оценкой генеральной доли $P(X \geq 124)$ является выборочная доля

$$w(X \geq 124) = (19+10+2)/100 = 0{,}31. \blacktriangleright$$

Оценка генеральной средней. Пусть из генеральной совокупности объема N отобрана случайная выборка X_1, X_2,...,X_k,...,X_n, где X_k — случайная величина, выражающая значение признака у k-го элемента выборки (k=1, 2,...,n). Следует найти «наилучшую» оценку для генеральной средней.

Рассмотрим в качестве такой возможной оценки выборочную среднюю[1] $\bar{x}$ (вспомним, что в примере **9.4** именно $\bar{x}$ явилась оценкой метода наименьших квадратов для $\bar{x}_0$), т.е.

$$\bar{x} = \frac{\sum_{k=1}^{n} X_k}{n}.$$

а) **Выборка повторная**

Закон распределения для каждой случайной величины X_k (k=1,2,..., n) имеет вид:

x_i	x_1	x_2	...	x_i	...	x_m
p_i	$\frac{N_1}{N}$	$\frac{N_2}{N}$	...	$\frac{N_i}{N}$	...	$\frac{N_m}{N}$

(9.13)

Действительно, вероятность того, что 1-й отобранный в выборку элемент имеет значение признака x_1, согласно классическому определению вероятности равна $p(X_1 = x_1) = \frac{N_1}{N}$, так как из общего числа N элементов генеральной совокупности N_1 элементов имеют значение признака x_1. Так как выборка повторная и каждый отобранный и обследованный элемент возвращается в исходную совокупность, восстанавливая всякий раз ее первоначальные состав и объем, то вероятность $p(X_k=x_1) = \frac{N_1}{N}$ для любого элемента выборки, т.е. для k=1,2,...,n. Аналогично

[1] См. замечание на с. 307.

можно определить $p(X_k=x_i)=\frac{N_i}{N}$ для $k=1,2,...,n$; $i=1,2,...,m$ и убедиться в том, что закон распределения каждой случайной величины X_k один и тот же — **(9.13)**.

Случайные величины $X_1,X_2,...,X_k,...,X_n$ *независимы*, так как независимы любые события $X_k=x_i$ ($k=1,2,...n$; $i=1,2,...,m$) и их комбинации. Например, независимы события $X_2=x_1$ и $X_1=x_1$, ибо $p_{X_1=x_1}(X_2=x_1)=p(X_2=x_1)=\frac{N_1}{N}$, т.е. вероятность того, что значение признака у 2-го отобранного в выборку элемента равно x_1, не меняется в зависимости от того, какое значение признака у 1-го элемента, и т.д.

Найдем числовые характеристики случайной величины X_k:

$$M(X_k)=\sum_{i=1}^{m}x_ip_i=\frac{\sum_{i=1}^{m}x_iN_i}{N}=\bar{x}_0, \qquad \textbf{(9.14)}$$

$$D(X_k)=\sum_{i=1}^{m}(x_i-\bar{x}_0)^2p_i=\frac{\sum_{i=1}^{m}(x_i-\bar{x}_0)^2N_i}{N}=\sigma^2, \qquad \textbf{(9.15)}$$

т.е. *математическое ожидание и дисперсия каждой случайной величины X_k — это соответственно генеральная средняя и генеральная дисперсия.*

Теорема. *Выборочная средняя $\bar{x}$ повторной выборки есть несмещенная и состоятельная оценка генеральной средней $\bar{x}_0$, причем*

$$\sigma_{\bar{x}}^2=\frac{\sigma^2}{n}. \qquad \textbf{(9.16)}$$

□ Докажем вначале несмещенность оценки. Найдем математическое ожидание выборочной средней $\bar{x}$, учитывая **(9.14)**:

$$M(\bar{x})=M\left(\frac{\sum_{k=1}^{n}X_k}{n}\right)=\frac{\sum_{k=1}^{n}M(X_k)}{n}=\frac{\sum_{k=1}^{n}\bar{x}_0}{n}=\frac{n\bar{x}_0}{n}=\bar{x}_0,$$

т.е. $\bar{x}$ — несмещенная оценка для $\bar{x}_0$.

Найдем дисперсию выборочной средней $\bar{x}$, учитывая **(9.15)** и то, что $X_1, X_2,...,X_k,...,X_n$ — независимые случайные величины:

$$\sigma_x^2=D(\bar{x})=D\left(\frac{\sum_{k=1}^{n}X_k}{n}\right)=\frac{1}{n^2}\sum_{k=1}^{n}D(X_k)=\frac{1}{n^2}\sum_{k=1}^{n}\sigma^2=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n}.$$

Осталось доказать состоятельность оценки $\bar{x}$, которая следует непосредственно из теоремы Чебышева **(6.14)**:

$$\lim_{n\to\infty}P(|\bar{x}-\bar{x}_0|\le\varepsilon)=1,$$

или $$\bar{x}\xrightarrow[n\to\infty]{\mathcal{P}}\bar{x}_0.$$

б) **Выборка бесповторная**

В этом случае случайные величины $X_1,X_2,...,X_n$ будут *зависимыми*. Рассмотрим, например, события $X_1=x_1$ и $X_2=x_1$.

Теперь вероятность $p_{X_1=x_1}(X_2=x_1)=\frac{N_1-1}{N-1}$, так как отобранный элемент (в случае бесповторной выборки) в исходную совокупность не возвращается, то в ней остается всего $N-1$ элементов, из которых со значением признака — N_1-1. Эта вероятность $p_{X_1=x_1}(X_2=x_1)$ не равна $p(X_2=x_1)=\frac{N_1}{N}$, т.е. события $X_1=x_1$ и $X_2=x_1$ — зависимые. Аналогично будут зависимыми любые события $X_k=x_i$ ($k=1,2,...,n$; $i=1,2,...,m$), а значит, зависимы случайные величины X_1, X_2,..., X_k,...,X_n.

Однако и для бесповторной выборки выборочная средняя является «хорошей» оценкой. Об этом свидетельствует теорема.

Теорема. *Выборочная средняя $\bar{x}$ бесповторной выборки есть несмещенная и состоятельная оценка генеральной средней $\bar{x}_0$, причем*

$$\sigma_{\bar{x}}'^2=\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\approx\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right) \qquad \textbf{(9.17)}$$

Теорему принимаем без доказательства. Частные случаи формулы **(9.17)** аналогичны **(9.12)** (см. с. 310).

▷ **Пример 9.6.** Найти несмещенную и состоятельную оценку средней выработки рабочих цеха по данным выборки, представленной в табл. 8.1.

Р е ш е н и е. Несмещенная и состоятельная оценка генеральной средней $\bar{x}_0$ есть выборочная средняя $\bar{x}$, найденная в примере **8.3**, т.е. $\bar{x}=119{,}2(\%)$. ▶

Оценка генеральной дисперсии. На первый взгляд, наиболее подходящей оценкой для генеральной дисперсии σ^2 является выборочная дисперсия s^2. Следующая теорема свидетельствует о том, что s^2 не является «наилучшей» оценкой.

Теорема. *Выборочная дисперсия s^2 повторной и бесповторной выборок есть смещенная и состоятельная оценка генеральной дисперсии σ^2.*

□ Принимая без доказательства состоятельность оценки s^2, докажем, что она — смещенная оценка. В соответствии с **(8.10)** $s^2=\overline{x^2}-\bar{x}^2$. На основании свойства **3** средней арифметической (§ 8.2) и дисперсии (§ 8.3), если все значения признака уменьшить на одно и то же число c, то средняя уменьшится на это число, т.е. $\overline{x-c}=\bar{x}-c$, а дисперсия не изменится:

$$s^2=s_x^2=s_{x-c}^2=\overline{(x-c)^2}-\overline{(x-c)}^2=\overline{(x-c)^2}-(\bar{x}-c)^2.$$

Полагая $c=\bar{x}_0$, получим

$$s^2=\overline{(x-\bar{x}_0)^2}-(\bar{x}-\bar{x}_0)^2.$$

а) **Выборка повторная**

Для повторной выборки выборочные значения рассматриваем как *независимые* случайные величины X_1, $X_2,\ldots,X_k,\ldots,X_n$, каждая из которых имеет один и тот же закон распределения **(9.13)** с числовыми характеристиками **(9.14)** и **(9.15)**, т.е. $M(X_k)=\bar{x}_0$, $D(X_k)=\sigma^2$, $k=1,2,\ldots,n$.

Найдем математическое ожидание оценки s^2:

$$M(s^2)=M\left(\frac{\sum_{k=1}^{n}(X_k-\bar{x}_0)^2}{n}\right)-M(\bar{x}-\bar{x}_0)^2.$$

Первый член в правой части

$$M\left(\frac{\sum_{k=1}^{n}(X_k-\bar{x}_0)^2}{n}\right)=\frac{\sum_{k=1}^{n}M(X_k-\bar{x}_0)^2}{n}=\frac{\sum_{k=1}^{n}D(X_k)}{n}=\frac{\sum_{k=1}^{n}\sigma^2}{n}=\frac{n\sigma^2}{n}=\sigma^2.$$

Второй член с учетом того, что $\bar{x}$ есть несмещенная оценка $\bar{x}_0$, т.е. $M(\bar{x})=\bar{x}_0$, и **(9.16)**,

$$M(\bar{x}-\bar{x}_0)^2=D(\bar{x})=\sigma_{\bar{x}}^2=\frac{\sigma^2}{n}.$$

Поэтому

$$M(s^2)=\sigma^2-\frac{\sigma^2}{n}=\frac{n-1}{n}\sigma^2.$$

б) **Выборка бесповторная**

Как уже рассмотрено выше, для бесповторной выборки X_1, $X_2,\ldots,X_n$ — *зависимые* случайные величины. Можно показать, что

$$M(s^2)=\frac{n-1}{n}\frac{N}{N-1}\sigma^2\approx\frac{n-1}{n}\sigma^2$$

(так как объем генеральной совокупности N, как правило, большой и $N\approx N-1$).

Итак, и для повторной выборки, и для бесповторной

$M(s^2)=\frac{n-1}{n}\sigma^2$, т.е. s^2 — смещенная[1] оценка σ^2. ■

Так как $\frac{n-1}{n}<1$ и $M(s^2)<\sigma^2$, то *выборочная дисперсия* (в среднем, полученная по разным выборкам) *занижает генеральную дисперсию*. Поэтому, заменяя σ^2 на s^2, мы допускаем систематическую погрешность в меньшую сторону. Чтобы ее ликвидировать, достаточно ввести поправку, умножив s^2 на $\frac{n}{n-1}$. Тогда с учетом **(9.4)** получим *«исправленную» выборочную дисперсию*

[1] Так как смещение оценки $b(s^2)=M(s^2)-\sigma^2=\frac{n-1}{n}\sigma^2-\sigma^2=-\frac{\sigma^2}{n}$ при $n\to\infty$ стремится к нулю, т.е. $\lim_{n\to\infty}b(s^2)=0$, то s^2 *есть асимптотически несмещенная оценка* σ^2.

$$\hat{s}^2 = \frac{n}{n-1}s^2 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2 n_i}{n-1}. \qquad (9.18)$$

Очевидно, что

$$M(\hat{s}^2) = M\left(\frac{n}{n-1}s^2\right) = \frac{n}{n-1}M(s^2) = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2.$$

т.е. $\hat{s}^2$ *является несмещенной и состоятельной оценкой генеральной дисперсии* σ^2.

▷ **Пример 9.7.** Найти несмещенную и состоятельную оценку дисперсии случайной величины X — выработки рабочих цеха по данным выборки, представленной в табл. 8.1.

Р е ш е н и е. Несмещенной и состоятельной оценкой дисперсии случайной величины X (генеральной дисперсии) σ^2 является «исправленная» выборочная дисперсия $\hat{s}^2$. В примере **8.6** вычислена выборочная дисперсия $s^2 = 87{,}48$. На основании **(9.18)** при n=100

$$\hat{s}^2 = \frac{100}{99}\cdot 87{,}48 = 88{,}36. \blacktriangleright$$

Разница между s^2 и $\hat{s}^2$ заметна при небольшом числе наблюдений n. При n>30–40 $\hat{s}^2 \approx s^2$, т.е. в качестве оценки для σ^2 вполне можно использовать выборочную дисперсию s^2.

9.5. Определение эффективных оценок с помощью неравенства Рао—Крамера—Фреше

Выше рассмотрены оценки параметров (характеристик) генеральной совокупности с точки зрения их состоятельности и несмещенности. Однако до сих пор не были затронуты вопросы э ф ф е к т и в н о с т и этих оценок.

Пусть $\varphi(x,\theta)$ — плотность вероятности признака X (случайной величины X — генеральной совокупности), если X непрерывна, и функция вероятностей $\varphi(x_i,\theta) = P(X = x_i,\theta)$, если X дискретна.

Для широкого класса генеральных совокупностей (при выполнении так называемых *условий регулярности* функции $\varphi(x,\theta)$: дифференцируемости по θ, независимости области определения от θ и т.д., являющихся достаточно общими) имеет место *неравенство Рао—Крамера—Фреше* (*неравенство информации*):

$$D(\tilde{\theta}_n) \geq \frac{1}{nI(\theta)}, \qquad (9.19)$$

где $D(\tilde{\theta}_n)$ — дисперсия оценки $\tilde{\theta}_n$ параметра θ;

$I(\theta)$ — *количество информации* Фишера о параметре θ, содержащееся в единичном наблюдении и определяемое в дискретном случае формулой:

$$I(\theta) = M\left[\left(\ln\varphi(X,\theta)\right)'_\theta\right]^2 = \sum_{i=1}^{m}\left[\frac{\varphi'_\theta(x_i,\theta)}{\varphi(x_i,\theta)}\right]^2 \varphi(x_i,\theta), \qquad (9.20)$$

а в непрерывном случае — формулой:

$$I(\theta) = M\left[\left(\ln\varphi(X,\theta)\right)'_\theta\right]^2 = \int_{-\infty}^{+\infty}\left[\frac{\varphi'_\theta(x,\theta)}{\varphi(x,\theta)}\right]^2 \varphi(x,\theta)\,d\theta. \qquad (9.21)$$

Неравенство информации позволяет найти тот минимум $\min D(\tilde{\theta}_n)$, который должна иметь дисперсия оценки $\sigma^2_{\tilde{\theta}_n}$, чтобы быть эффективной оценкой $\tilde{\theta}^э_n$, т.е. $\sigma^2_{\tilde{\theta}^э_n} = \min D(\tilde{\theta}_n)$.

▷ **Пример 9.8.** Найти эффективную оценку генеральной доли p повторной выборки.

Р е ш е н и е. Найдем количество информации Фишера $I(p)$ по формуле **(9.20)**. Напомним (см. § 9.4), что в данном случае наблюдаемая величина X принимает два значения — 0 и 1 с вероятностями соответственно: $\varphi(0;p) = q = 1-p$ и $\varphi(1;p) = p$. Имеем

$$I(p) = \left[\frac{\varphi'_p(0;p)}{\varphi(0;p)}\right]^2 \varphi(0;p) + \left[\frac{\varphi'_p(1;p)}{\varphi(1;p)}\right]^2 \varphi(1;p) =$$

$$= \left(\frac{-1}{1-p}\right)^2 (1-p) + \left(\frac{1}{p}\right)^2 \cdot p = \frac{1}{1-p} + \frac{1}{p} = \frac{1}{p(1-p)},$$

т.е.
$$\min D(\tilde{\theta}_n) = \frac{1}{nI(p)} = \frac{p(1-p)}{n}.$$

Как показано выше, именно такую дисперсию (см. **(9.11)**) имеет дисперсия выборочной доли w повторной выборки:

$\sigma_w^2 = \dfrac{pq}{n} = \dfrac{p(1-p)}{n}$. Следовательно, *выборочная доля w повторной выборки есть эффективная оценка генеральной доли p.* ▶

▷ **Пример 9.9.** Найти эффективную оценку генеральной средней x_0 (математического ожидания a) повторной выборки для нормально распределенной генеральной совокупности.

Р е ш е н и е. В случае нормального закона распределения плотность вероятности

$$\varphi_N(x,a) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Тогда $$\ln \varphi_N(x,a) = -\ln\sqrt{2\pi\sigma^2} - \frac{(x-a)^2}{2\sigma^2}$$

и $$\left[\left(\ln \varphi_N(x,a)\right)'_a\right]^2 = \left(0 + \frac{x-a}{\sigma^2}\right)^2 = \frac{(x-a)^2}{\sigma^4}.$$

Теперь количество информации Фишера

$$I(a) = M\left[\left(\ln \varphi_N(X,a)\right)'_a\right]^2 = M\left[\frac{(X-a)^2}{\sigma^4}\right] = \frac{D(X)}{\sigma^4} = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

Минимально возможная оценка дисперсии оценки

$$\min D(\tilde{\theta}_n) = \frac{1}{nI(a)} = \frac{\sigma^2}{n}.$$

Выше (см. § 9.4) мы установили, что именно такую дисперсию (см. **(9.16)**) имеет выборочная средняя $\bar{x}$ повторной выборки: $\sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{n}$. Итак, *выборочная средняя $\bar{x}$ повторной выборки для нормально распределенной генеральной совокупности является эффективной оценкой генеральной средней $\bar{x}_0$.* ▶

Аналогично примеру **9.9** можно показать, что эффективная оценка $\tilde{\theta}_n^3$ генеральной дисперсии σ^2 повторной выборки для нормально распределенной генеральной совокупности должна иметь минимальную дисперсию $\min D(\tilde{\theta}_n) = \dfrac{2\sigma^4}{n}$.

В то же время дисперсия исправленной выборочной дисперсии $\hat{s}^2$, являющейся несмещенной оценкой генеральной дисперсии σ^2, как можно показать, есть $\sigma^2_{\hat{s}^2_n} = \dfrac{2\sigma^4}{n-1}$, т.е. *исправленная выборочная дисперсия $\hat{s}^2$ повторной выборки не является эффективной оценкой генеральной дисперсии σ^2.*

Выборочные дисперсии — s^2 и «исправленная» $\hat{s}^2$ — являются асимптотически эффективными оценками генеральной дисперсии σ^2, так как при $n \to \infty$ их эффективности, вычисленные по формуле **(9.8)**, стремятся к единице.

Эффективной же оценкой генеральной дисперсии σ^2 является статистика

$$s_*^2 = \frac{1}{n}\sum_{i=1}^{m}(x_i - \bar{x}_0)^2 n_i, \qquad \textbf{(9.22)}$$

но для ее нахождения надо знать генеральную среднюю $\bar{x}_0$, которая в большинстве случаев применения выборочного метода неизвестна.

В заключение отметим, что не для всякого закона распределения может быть использовано неравенство Рао—Крамера—Фреше для нахождения эффективных оценок параметров, поскольку не всегда оказываются выполнены условия регулярности функции $\varphi(x,\theta)$. Так, например, с помощью неравенства информации нельзя получить эффективные оценки для параметров a и b равномерного закона распределения, так как они непосредственно задают границы области определения функции $\varphi(x,\theta)$.

9.6. Понятие интервального оценивания. Доверительная вероятность и предельная ошибка выборки

Выше рассмотрена оценка параметров θ генеральной совокупности о д н и м ч и с л о м, т.е. $\bar{x}_0$ — числом $\bar{x}$, p — числом w, σ^2 — числом s^2 или $\hat{s}^2$. Такие оценки параметров называются *точечными.*

Однако точечная оценка $\tilde{\theta}_n$ является лишь приближенным значением неизвестного параметра θ даже в том случае, если она несмещенная (в среднем совпадает с θ), состоятельная (стремится к θ с ростом n) и эффективная (обладает наименьшей степенью случайных отклонений от θ) и для выборки малого объема может существенно отличаться от θ.

Чтобы получить представление о точности и надежности оценки $\tilde{\theta}_n$ параметра θ, используют интервальную оценку параметра.

Определение. ***Интервальной оценкой параметра*** θ *называется числовой интервал* $\left(\tilde{\theta}_n^{(1)}, \tilde{\theta}_n^{(2)}\right)$, *который с заданной вероятностью* γ *накрывает неизвестное значение параметра* θ (рис. 9.1).

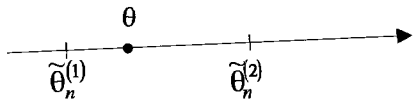


Рис. 9.1

Обращаем внимание на то, что границы интервала $\left(\tilde{\theta}_n^{(1)}, \tilde{\theta}_n^{(2)}\right)$ и его величина находятся по выборочным данным и потому являются случайными величинами в отличие от оцениваемого параметра θ — величины неслучайной, поэтому правильнее говорить о том, что интервал $\left(\tilde{\theta}_n^{(1)}, \tilde{\theta}_n^{(2)}\right)$ «накрывает», а не «содержит» значение θ.

Такой *интервал* $\left(\tilde{\theta}_n^{(1)}, \tilde{\theta}_n^{(2)}\right)$ *называется* ***доверительным***, *а вероятность* γ — ***доверительной вероятностью***, ***уровнем доверия*** или ***надежностью оценки***.

Величина доверительного интервала существенно зависит от объема выборки n (уменьшается с ростом n) и от значения доверительной вероятности γ (увеличивается с приближением γ к единице).

Очень часто (но не всегда) доверительный интервал выбирается симметричным относительно параметра θ, т.е. $(\theta-\Delta,\ \theta+\Delta)$.

Наибольшее отклонение Δ оценки $\tilde{\theta}_n$ от оцениваемого параметра θ, в частности, *выборочной средней* (*или доли*) *от генеральной средней* (*или доли*), *которое возможно с заданной доверительной вероятностью* γ, *называется* ***предельной ошибкой выборки***.

Ошибка Δ *является* ***ошибкой репрезентативности*** (***представительства***) *выборки. Она возникает только вследствие того, что исследуется не вся совокупность, а лишь часть ее* (*выборка*), *отобранная случайно.* Эту ошибку часто называют *случайной* ошибкой репрезентативности. Ее не следует путать с *систематической* ошибкой репрезентативности, появляющейся в результате нарушения принципа случайности при отборе элементов в выборку.

Построение доверительного интервала для генеральной средней и генеральной доли по большим выборкам. Для построения доверительных интервалов для параметров генеральных совокупностей могут быть реализованы два подхода, основанных на знании *точного* (при данном объеме выборки n) или *асимптотического* (при $n \to \infty$) распределения выборочных характеристик (или некоторых функций от них). Первый подход реализован далее при построении интервальных оценок параметров для малых выборок (см. § 9.7). В данном параграфе рассматривается второй подход, применимый для больших выборок (порядка сотен наблюдений).

Теорема. *Вероятность того, что отклонение выборочной средней* (*или доли*) *от генеральной средней* (*или доли*) *не превзойдет числа* $\Delta > 0$ (*по абсолютной величине*), *равна*:

$$P\left(|\bar{x}-\bar{x}_0| \le \Delta\right) = \Phi(t) = \gamma, \quad \textbf{(9.23)}$$

где $t = \dfrac{\Delta}{\sigma_x}$,

$$P\left(|w-p| \le \Delta\right) = \Phi(t) = \gamma, \quad \textbf{(9.24)}$$

где $t = \dfrac{\Delta}{\sigma_w}$,

$\Phi(t)$ — функция (интеграл вероятностей) Лапласа.

□ Выше (§ 9.4) показано, что выборочная средняя $\bar{x}$ и выборочная доля w повторной выборки представляют сумму n независимых случайных величин $\dfrac{\sum_{k=1}^{n} X_k}{n} = \sum_{k=1}^{n} \dfrac{X_k}{n}$, где X_k (k=1,2,...,n) имеет один и тот же закон распределения — соответственно **(9.13)** и **(9.10)** с конечными математическим ожиданием и дисперсией. Следовательно, на основании теоремы Ляпунова (см. § 6.5) при $n \to \infty$ распределения $\bar{x}$ и w неограниченно приближаются к нормальным (практически при n>30—40 распределения $\bar{x}$ и w можно считать приближенно нормальными).

Для бесповторной выборки $\bar{x}$ и w представляют сумму зависимых случайных величин. Однако можно показать, что и в этом случае при $n \to \infty$ закон распределения $\bar{x}$ и w как угодно близко приближается к нормальному.

Формулы **(9.23)** и **(9.24)** следуют непосредственно из свойства 2 нормального закона (см. § 4.7, формулы **(4.34)**, **(4.35)**). ■

Формулы **(9.23)** и **(9.24)** получили название *формул доверительной вероятности для средней и доли.*

Определение. *Среднее квадратическое отклонение выборочной средней* σ_x *и выборочной доли* σ_w *собственно-случайной выборки называется* ***средней квадратической (стандартной) ошибкой выборки.*** (Для бесповторной выборки обозначаем соответственно $\sigma'_{\bar{x}}$ и σ'_w).

Из рассмотренной теоремы вытекают следующие следствия.

Следствие 1. *При заданной доверительной вероятности* γ *предельная ошибка выборки равна t-кратной величине средней квадратической ошибки, где* $\Phi(t)=\gamma$, т.е.

$$\Delta = t\sigma_{\bar{x}}, \qquad (9.25)$$

$$\Delta = t\sigma_w. \qquad (9.26)$$

Следствие 2. *Интервальные оценки (доверительные интервалы) для генеральной средней и генеральной доли могут быть найдены по формулам:*

$$\bar{x} - \Delta \le \bar{x}_0 \le \bar{x} + \Delta, \qquad (9.27)$$

$$w - \Delta \le p \le w + \Delta. \qquad (9.28)$$

Формулы средних квадратических ошибок выборки $\sigma_{\bar{x}}$, $\sigma'_{\bar{x}}$, σ_w, σ'_w могут быть легко получены из формул **(9.16)**, **(9.17)**, **(9.11)**, **(9.12)** соответствующих дисперсий $\sigma^2_{\bar{x}}$, $\sigma'^2_{\bar{x}}$, σ^2_w, σ'^2_w. Поместим их в таблицу:

Таблица 9.2

Оцениваемый параметр	*Формулы средних квадратических ошибок выборки*	
	повторная выборка	*бесповторная выборка*
Средняя	$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} \approx \sqrt{\frac{s^2}{n}}$ **(9.29)**	$\sigma'_{\bar{x}} \approx \sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)}$ **(9.30)**
Доля	$\sigma_w = \sqrt{\frac{pq}{n}} \approx \sqrt{\frac{w(1-w)}{n}}$ **(9.31)**	$\sigma'_w \approx \sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}$ **(9.32)**

Так как генеральные доля p и дисперсия[1] σ^2 неизвестны, то в формулах табл. 9.2 заменяем их состоятельными оценками по выборке — соответственно w и s^2, ибо при достаточно большом объеме выборки n практически достоверно, что $w \approx p$, $s^2 \approx \sigma^2$. При определении средней квадратической ошибки выборки для доли, если даже w неизвестна, в качестве pq можно взять его максимально возможное значение

$$(pq)_{\max} = [p(1-p)]_{\max} = 0{,}5 \cdot 0{,}5 = 0{,}25$$

(так как $pq = p(1-p) = -(p^2 - p) = 0{,}25 - (p - 0{,}5)^2$, то pq максимально при $p = 0{,}5$).

▷ **Пример 9.10.** При обследовании выработки 1000 рабочих цеха в отчетном году по сравнению с предыдущим по схеме собственно-случайной выборки было отобрано 100 рабочих. Получены следующие данные (см. первые две графы табл. 8.1). Необходимо определить: а) вероятность того, что средняя выработка рабочих цеха отличается от средней выборочной не более, чем на 1% (по абсолютной величине); б) границы, в которых с вероятностью 0,9545 заключена средняя выработка рабочих цеха. Рассмотреть случаи повторной и бесповторной выборки.

Р е ш е н и е. а) Имеем $N = 1000$, $n = 100$. Ранее в примере **8.8** были вычислены $x = 119{,}2(\%)$, $s^2 = 87{,}48$.

а) Найдем среднюю квадратическую ошибку выборки для средней:

для повторной выборки

По **(9.29)**

$$\sigma_{\bar{x}} = \sqrt{\frac{87{,}48}{100}} = 0{,}935(\%).$$

для бесповторной выборки

По **(9.30)**

$$\sigma'_{\bar{x}} = \sqrt{\frac{87{,}48}{100}\left(1-\frac{100}{1000}\right)} = 0{,}887(\%).$$

[1] Заметим, что в формуле **(9.29)** σ^2 представляет дисперсию *количественного* признака X (генеральной совокупности), а в формуле **(9.31)** величина $pq = p(1-p)$ — дисперсию *альтернативного* признака X.

Теперь искомую доверительную вероятность находим по (**9.23**):

$$P\left(\left|\bar{x}-\bar{x}_0\right|\le 1\right)= \qquad P\left(\left|\bar{x}-\bar{x}_0\right|\le 1\right)=$$

$$=\Phi\left(\frac{1}{0{,}935}\right)=\Phi(1{,}07)=0{,}715. \qquad =\Phi\left(\frac{1}{0{,}887}\right)=\Phi(1{,}13)=0{,}741.$$

(Значения $\Phi(t)$ находим по табл. II приложений.)

Итак, вероятность того, что выборочная средняя отличается от генеральной средней не более чем на 1% (по абсолютной величине), равна 0,715 для повторной и 0,741 для бесповторной выборки.

б) Найдем предельные ошибки повторной и бесповторной выборок по формуле (**9.25**), в которой $t = 2{,}00$ (находим по табл. II приложений при данной в условии доверительной вероятности γ из соотношения $\gamma=\Phi(t)=0{,}9545$).

$$\Delta=2{,}00\cdot 0{,}935=1{,}870(\%). \qquad \Delta'=2{,}00\cdot 0{,}887=1{,}774(\%).$$

Теперь искомый доверительный интервал определяем по (**9.27**):

$$119{,}2-1{,}870\le\bar{x}_0\le 119{,}2+1{,}870 \qquad 119{,}2-1{,}774\le\bar{x}_0\le 119{,}2+1{,}774$$

$$\text{или } 117{,}33\le\bar{x}_0\le 121{,}07(\%). \qquad \text{или } 117{,}43\le\bar{x}_0\le 120{,}97(\%).$$

Таким образом, с надежностью 0,9545 средняя выработка рабочих цеха заключена в границах от 117,33 до 121,07%, если выборка повторная, и от 117,43 до 120,97%, если выборка бесповторная. ▶

▷ **Пример 9.11.** Из партии, содержащей 2000 деталей, для проверки по схеме собственно-случайной бесповторной выборки было отобрано 200 деталей, среди которых оказалось 184 стандартных. Найти: а) вероятность того, что доля нестандартных деталей во всей партии отличается от полученной доли в выборке не более чем на 0,02 (по абсолютной величине); б) границы, в которых с надежностью 0,95 заключена доля нестандартных деталей во всей партии.

Р е ш е н и е. Имеем $N = 2000$, $n = 200$, $m = 200—184 = 16$ нестандартных деталей. Выборочная доля нестандартных деталей $w=\frac{m}{n}=\frac{16}{200}=0{,}08.$

а) По (**9.32**) найдем среднюю квадратическую ошибку бесповторной выборки для доли:

$$\sigma'_w=\sqrt{\frac{0{,}08\cdot 0{,}92}{200}\left(1-\frac{200}{2000}\right)}=0{,}0182.$$

Теперь искомую доверительную вероятность находим по (**9.24**):

$$P\left(\left|w-p\right|\le 0{,}02\right)=\Phi\left(\frac{0{,}02}{0{,}0182}\right)=$$

$$=\Phi(1{,}10)=0{,}729 \text{ (по табл. II приложений)},$$

т.е. вероятность того, что выборочная доля нестандартных деталей будет отличаться от генеральной доли не более чем на 0,02 (по абсолютной величине), равна 0,729.

б) Учитывая, что $\gamma=\Phi(t)=0{,}95$ и (по таблице) $t = 1{,}96$, найдем предельную ошибку выборки для доли по (**9.26**): $\Delta=1{,}96\cdot 0{,}0182=0{,}0357$. Теперь искомый доверительный интервал определяем по (**9.28**): $0{,}08-0{,}0357\le p\le 0{,}08+0{,}0357$ или $0{,}044\le p\le 0{,}116$.

Итак, с надежностью 0,95 доля нестандартных деталей во всей партии заключена от 0,044 до 0,116. ▶

Объем выборки. Для проведения выборочного наблюдения весьма важно правильно установить объем выборки n, который в значительной степени определяет необходимые при этом временные, трудовые и стоимостные затраты. *Для определения n необходимо задать надежность (доверительную вероятность) оценки* γ *и точность (предельную ошибку выборки)* Δ.

Объем выборки находится из формулы, выражающей предельную ошибку выборки через дисперсию признака. Например, для повторной выборки при оценке генеральной средней с надежностью γ с учетом (**9.25**) и (**9.29**) эта формула имеет вид: $\Delta=t\sqrt{\frac{\sigma^2}{n}}$, откуда $n=\frac{t^2\sigma^2}{\Delta^2}$, где $\Phi(t)=\gamma$. Аналогично могут быть получены и другие формулы объема выборки, которые сведем в таблицу:

Таблица 9.3

Оцениваемый параметр	*Повторная выборка*		*Бесповторная выборка*	
Генеральная средняя	$n=\dfrac{t^2\sigma^2}{\Delta^2}$	**(9.33)**	$n'=\dfrac{Nt^2\sigma^2}{t^2\sigma^2+N\Delta^2}$	**(9.34)**
Генеральная доля	$n=\dfrac{t^2pq}{\Delta^2}$	**(9.35)**	$n'=\dfrac{Nt^2pq}{t^2pq+N\Delta^2}$	**(9.36)**

Если найден объем повторной выборки n, то объем соответствующей бесповторной выборки n' можно определить по формуле:

$$n'=\frac{nN}{n+N}. \qquad \textbf{(9.37)}$$

Так как $\dfrac{N}{n+N}<1$, то *при одних и тех же точности и надежности оценок объем бесповторной выборки n' всегда меньше объема повторной выборки n.* Этим и объясняется тот факт, что на практике в основном используется бесповторная выборка.

Как видно из формул **(9.33)**—**(9.36)**, для определения объема выборки необходимо знать характеристики генеральной совокупности σ^2 или p, которые неизвестны и для определения которых предполагается провести выборочное наблюдение. В качестве этих характеристик обычно используют выборочные данные s^2 или w предшествующего исследования в аналогичных условиях, т.е. полагают $\sigma^2\approx s^2$ (или $\hat{s}^2$) или $p\approx w$.

Если никаких сведений о значениях σ^2 или p нет, то организуют специальную п р о б н у ю выборку небольшого объема, находят оценку $\hat{s}^2$ (более точную, чем s^2 для малой выборки) или w и, полагая $\sigma^2\approx\hat{s}^2$ или $p\approx w$, находят объем «о с н о в н о й» выборки.

При оценке генеральной доли (если о ней ничего неизвестно) вместо проведения пробной выборки можно в формулах **(9.35)**, **(9.36)** в качестве $pq = p(1 - p)$ взять его максимально возможное значение, равное 0,25, но при этом надо учитывать, что найденное значение объема выборки будет больше (иногда существенно больше) минимально необходимого для заданных точности и надежности оценок.

▷ **Пример 9.12.** По условию примера **9.10** определить объем выборки, при котором с вероятностью 0,9973 отклонение средней выработки рабочих в выборке от средней выработки всех рабочих цеха не превзойдет 1% (по абсолютной величине).

Р е ш е н и е. В качестве неизвестного значения σ^2 для определения объема выборки берем его состоятельную оценку $s^2 = 87{,}48$, найденную ранее в примере **9.10**.

Учитывая, что $\gamma=\Phi(t)=0{,}9973$ и (по табл. II приложений) $t = 3{,}00$, найдем объем повторной выборки по **(9.33)**, т.е. $n = 3^2\cdot 87{,}48/1 = 787$.

Объем бесповторной выборки по **(9.34)**:

$$n'=\frac{1000\cdot 3^2\cdot 87{,}48}{3^2\cdot 87{,}48+1000\cdot 1}=440{,}5\approx 441.$$

Объем бесповторной выборки n' мог быть вычислен и по **(9.37)**, так как уже известен объем повторной выборки n, т.е.

$$n'=\frac{787\cdot 1000}{787+1000}\approx 441.$$

Как видим, при одной и той же точности $\Delta = 1(\%)$ и надежности $\gamma = 0{,}9973$ оценки объем бесповторной выборки существенно меньше, чем повторной. ▶

▷ **Пример 9.13.** По условию примера **9.11** определить число деталей, которые надо отобрать в выборку, чтобы с вероятностью 0,95 доля нестандартных деталей в выборке отличалась от генеральной доли не более, чем на 0,04 (по абсолютной величине). Найти то же число, если о доле нестандартных деталей, даже приблизительно, ничего неизвестно.

Р е ш е н и е. В качестве неизвестного значения генеральной доли p возьмем ее состоятельную оценку $w = 0{,}08$, найденную ранее в примере **9.11**.

Учитывая, что $\gamma=\Phi(t)=0{,}95$ и (по таблице) $t = 1{,}96$, найдем объем бесповторной выборки по **(9.36)**, т.е.

$$n'=\frac{2000\cdot 1{,}96^2\cdot 0{,}08\cdot 0{,}92}{1{,}96^2\cdot 0{,}08\cdot 0{,}92+2000\cdot 0{,}04^2}=162.$$

Если о доле p ничего, даже приблизительно, неизвестно, в формуле **(9.36)** полагаем $pq=(pq)_{\max} = 0{,}25$. Тогда

$$n'=\frac{2000\cdot 1{,}96^2\cdot 0{,}25}{1{,}96^2\cdot 0{,}25+2000\cdot 0{,}04^2}=462,$$

т.е. полученное возможное значение объема выборки оказалось существенно выше необходимого. ▶

Замечание. Если генеральная совокупность бесконечна $(N=\infty)$, либо объем бесповторной выборки значительно меньше объема генеральной совокупности $(n \ll N)$, расчеты средних квадратических ошибок (для средней и доли) и необходимого объема бесповторной выборки следует проводить по соответствующим формулам для повторной выборки.

Построение доверительного интервала для генеральной доли по умеренно большим выборкам. Объем выборки может быть не настолько велик (например, десятки наблюдений), чтобы использовать приближенную формулу **(9.31)** $\sigma_w \approx \sqrt{\frac{w(1-w)}{n}}$ вместо точной $\sigma_w = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$. В то же время распределение выборочной доли w можно по-прежнему считать приближенно нормальным. В этом случае, учитывая **(9.24)**, **(9.26)**, доверительный интервал для генеральной доли p следует искать из условия

$$|w-p| \le t\sqrt{\frac{p(1-p)}{n}}. \qquad \textbf{(9.38)}$$

Возводя обе части неравенства **(9.38)** в квадрат, преобразуем его к равносильному:

$$(w-p)^2 \le \frac{t^2}{n}p(1-p). \qquad \textbf{(9.39)}$$

Областью решения неравенства **(9.39)** является внутренняя часть эллипса, проходящего через точки (0;0) и (1;1) и имеющего в этих точках касательные, параллельные оси абсцисс.

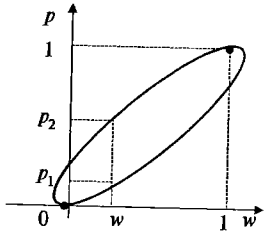


Рис. 9.2

Так как величина w заключена между 0 и 1, то область D нужно еще ограничить слева и справа прямыми $w=0$ и $w=1$ (наличие «лишних» областей, выходящих за полосу $0 \le w \le 1$, объясняется тем, что при значениях p, близких к 0 или 1, допущение о нормальном законе распределения w становится неправомерным).

По найденному по выборке значению w границы доверительного интервала (p_1, p_2)

для p определяются как точки пересечения соответствующей вертикальной прямой с эллипсом (рис. 9.2). Чем больше объем выборки n, тем «доверительный эллипс» более вытянут, тем уже доверительный интервал.

Границы p_1 и p_2 доверительного интервала для p могут быть найдены из соотношения **(9.39)** по формуле:

$$p_{1,2} = \frac{1}{1+t^2/n}\left[w + \frac{t^2}{2n} \mp t\sqrt{\frac{w(1-w)}{n} + \left(\frac{t}{2n}\right)^2}\right]. \qquad \textbf{(9.40)}$$

В случае больших выборок, при $n \to \infty$, величинами t^2/n (по сравнению с 1), $t^2/2n$ (по сравнению с w), $(t/2n)^2$ (по сравнению с $w(1-w)/n$) можно пренебречь, и получим:

$$p_{1,2} \approx w \mp t\sqrt{\frac{w(1-w)}{n}} \approx w \mp t\sigma_w = w \mp \Delta,$$

т.е. доказанные ранее формулы **(9.28)** и **(9.26)**.

▷ **Пример 9.14.** По данным примера **9.11** найти границы, в которых с надежностью 0,95 заключена доля p нестандартных изделий во всей партии, полагая $n = 50$, $w = 0{,}08$, $N=\infty$.

Решение. По формуле **(9.40)**, учитывая, что $t = 1{,}96$, найдем доверительные границы для генеральной доли p:

$$p_{1,2} = \frac{1}{1+1{,}96^2/50}\left[0{,}08 + \frac{1{,}96^2}{2\cdot 50} \mp 1{,}96\sqrt{\frac{0{,}08\cdot 0{,}92}{50} + \left(\frac{1{,}96}{2\cdot 50}\right)^2}\right] =$$

$$= 0{,}110 \mp 0{,}078 \text{ или } p_1 = 0{,}032,\ p_2 = 0{,}188,$$

т.е. с надежностью 0,95 доля нестандартных изделий во всей партии заключена от 0,032 до 0,188. ▶

9.7. Оценка характеристик генеральной совокупности по малой выборке

На практике часто приходится иметь дело с выборками небольшого объема n<10—20. В этом случае приведенный выше приближенный метод построения интервальной оценки для ге-

неральной средней и генеральной доли неприменим в силу двух обстоятельств:

1) необоснованным становится вывод о нормальном законе распределения выборочных средней $\bar{x}$ и доли w, так как он основан на центральной предельной теореме при больших n;

2) необоснованной становится замена неизвестных генеральной дисперсии σ^2 и доли p их точечными оценками соответственно s^2 (или $\hat{s}^2$) и w, так как в силу закона больших чисел (состоятельности оценок) эта замена возможна лишь при больших n.

Построение доверительного интервала для генеральной средней по малой выборке. Задача построения доверительного интервала для генеральной средней может быть решена, если в генеральной совокупности рассматриваемый признак имеет *нормальное распределение.*

Теорема. *Если признак (случайная величина) X имеет нормальный закон распределения с параметрами $M(X)=x_0$, $\sigma_x^2=\sigma^2$, т.е. $N(\bar{x}_0,\sigma^2)$, то выборочная средняя x при любом n (а не только при $n\to\infty$) имеет нормальный закон распределения $N\left(\bar{x}_0,\dfrac{\sigma^2}{n}\right)$.*

□ Если в случае больших выборок (при $n\to\infty$) из любых генеральных совокупностей нормальность распределения $x=\dfrac{\sum_{k=1}^{n}X_k}{n}$ обусловливалась суммированием большого числа одинаково распределенных случайных величин X_k/n (теорема Ляпунова), то в случае малых выборок, полученных из нормальной генеральной совокупности, нормальность распределения $\bar{x}$ вытекает из того, что распределение суммы (композиция) л ю б о г о ч и с л а нормально распределенных случайных величин имеет нормальное распределение (см. § 5.8). Формулы числовых характеристик для $\bar{x}\left(M(\bar{x})=\bar{x}_0, D(\bar{x})=\dfrac{\sigma^2}{n}\right)$ получены ранее (см. § 9.4, теорему на с. 312). ■

Таким образом, если бы была известна генеральная дисперсия σ^2, то доверительный интервал можно было бы построить анало-

гично изложенному выше и при малых n. Заметим, что в этом случае *нормированное отклонение выборочной средней* $t=\dfrac{\bar{x}-M(\bar{x})}{\sigma_{\bar{x}}}=\dfrac{\bar{x}-\bar{x}_0}{\sigma}\sqrt{n}$ *имеет стандартное нормальное распределение* $N(0,1)$, т.е. нормальное распределение с математическим ожиданием, равным нулю, и дисперсией, равной единице.

Действительно, используя свойства математического ожидания и дисперсии, получим, что

$$M(t)=M\left(\frac{\bar{x}-\bar{x}_0}{\sigma}\sqrt{n}\right)=\frac{\sqrt{n}}{\sigma}\left[M(\bar{x})-M(\bar{x}_0)\right]=\frac{\sqrt{n}}{\sigma}(\bar{x}_0-\bar{x}_0)=0,$$

$$\sigma_t^2=D(t)=D\left(\frac{\bar{x}-\bar{x}_0}{\sigma}\sqrt{n}\right)=\left(\frac{\sqrt{n}}{\sigma}\right)^2\left[D(\bar{x})+D(\bar{x}_0)\right]=\frac{n}{\sigma^2}\left(\frac{\sigma^2}{n}+0\right)=1.$$

Однако на практике почти всегда генеральная дисперсия σ^2 (как и оцениваемая генеральная средняя $\bar{x}_0$) неизвестна. Если заменить σ^2 ее «наилучшей» оценкой по выборке, а именно «исправленной» выборочной дисперсией $\hat{s}^2$, то большой интерес представляет распределение выборочной характеристики (статистики) $t=\dfrac{\bar{x}-\bar{x}_0}{\hat{s}}\sqrt{n}$ или, что то же с учетом **(9.18)**, распределение статистики $t=\dfrac{\bar{x}-\bar{x}_0}{s}\sqrt{n-1}$.

Представим статистику t в виде:

$$t=\frac{(\bar{x}-\bar{x}_0)/\dfrac{\sigma}{\sqrt{n}}}{\sqrt{\dfrac{1}{n-1}\dfrac{ns^2}{\sigma^2}}}. \qquad \textbf{(9.41)}$$

Числитель выражения **(9.41)**, как показано выше, имеет стандартное нормальное распределение $N(0;1)$. Можно показать (см. например, [3]), что случайная величина ns^2/σ^2 имеет χ^2-распределение с $k=n-1$ степенями свободы. Следовательно (см. § 4.9,

определение (**4.39**)), статистика t имеет t-распределение Стьюдента с $k = n - 1$ степенями свободы. Указанное распределение не зависит от неизвестных параметров распределения случайной величины X, а зависит лишь от числа k, называемого *числом степеней свободы*.

Выше (см. § 4.9) отмечено, что t-распределение Стьюдента напоминает нормальное распределение (см. рис. 4.17), и действительно при $k \to \infty$ как угодно близко приближается к нему.

Число степеней свободы k определяется как общее число n наблюдений (вариантов) случайной величины X минус число уравнений l, связывающих эти наблюдения, т.е. $k = n - l$.

Так, например, для распределения статистики $t = \dfrac{\bar{x} - \bar{x}_0}{s}\sqrt{n-1}$ число степеней свободы $k = n - 1$, ибо одна степень свободы «теряется» при определении выборочной средней $\bar{x}$ (n наблюдений связаны одним уравнением $\sum_{i=1}^{m} x_i n_i / n = \bar{x}$).

Зная t-распределение Стьюдента, можно найти такое критическое значение $t_{\gamma,n-1}$, что вероятность того, что статистика $t = \dfrac{\bar{x} - \bar{x}_0}{s}\sqrt{n-1}$ не превзойдет величину $t_{\gamma,n-1}$ (по абсолютной величине), равна γ:

$$P\left(\left|\frac{\bar{x} - \bar{x}_0}{s}\sqrt{n-1}\right| \le t_{\gamma,n-1}\right) = \theta(t, n-1) = \gamma. \qquad (9.42)$$

Функция $\theta(t,k) = 2\int_0^t \varphi(x,k)\,dx$, где $\varphi(x,k)$ — плотность вероятности t-распределения Стьюдента при числе степеней свободы k, табулирована. Эта функция аналогична функции Лапласа $\Phi(t)$, но в отличие от нее является функцией двух переменных — t и $k = n-1$. При $k \to \infty$ функция $\theta(t,k)$ неограниченно приближается к функции Лапласа $\Phi(t)$.

Формула доверительной вероятности (**9.42**) для малой выборки может быть представлена в равносильном виде:

$$P\left(|\bar{x} - \bar{x}_0| \le \Delta_{\text{м в}}\right) = \theta(t, n-1) = \gamma, \qquad (9.43)$$

где

$$\Delta_{\text{м.в}} = \frac{t_{\gamma,n-1}\, s}{\sqrt{n-1}} \qquad (9.44)$$

— предельная ошибка малой выборки. Доверительный интервал для генеральной средней, как и ранее, находится по формуле:

$$\bar{x} - \Delta_{\text{м.в}} \le \bar{x}_0 \le \bar{x} + \Delta_{\text{м.в}}. \qquad (9.45)$$

▷ **Пример 9.15.** Для контроля срока службы электроламп из большой партии было отобрано 17 электроламп. В результате испытаний оказалось, что средний срок службы отобранных ламп равен 980 ч, а среднее квадратическое отклонение их срока службы — 18 ч. Необходимо определить: а) вероятность того, что средний срок службы ламп во всей партии отличается от среднего срока службы отобранных для испытаний ламп не более чем на 8 ч (по абсолютной величине); б) границы, в которых с вероятностью 0,95 заключен средний срок службы ламп во всей партии.

Р е ш е н и е. Имеем по условию $n = 20$, $\bar{x} = 980$(ч), $s = 18$ ч.

а) Зная предельную ошибку малой выборки $\Delta_{\text{м в}} = 8$ (ч), найдем $t_{\gamma,n-1}$ из соотношения (**9.44**):

$$t_{\gamma,n-1} = \frac{\Delta_{\text{м.в}}}{s}\sqrt{n-1} = \frac{8}{18}\sqrt{17-1} = 1{,}78.$$

Теперь искомая доверительная вероятность по (**9.43**):

$P\left(|\bar{x} - \bar{x}_0| \le 8\right) = \theta(1{,}78;\, 16) = 0{,}906$, ($\theta(1{,}78;\, 16)$ находим по таблице значений[1] $\theta(t;k)$ при числе степеней свободы $k = 16$).

---

[1] Так как непосредственно таблица значений $\theta(t,k)$ в учебнике не приводится, то вероятность γ можно найти приближенно, используя табл. IV приложений, в которой указаны значения $t_{\gamma,k}$, полученные из условия $\theta(t_{\gamma,k}) = \gamma$. Так, для k=16 по этой таблице $\gamma = 0{,}9$ при $t = 1{,}75$. Более точно вероятность γ, соответствующую $t = 1{,}78$, можно найти, прибегнув к интерполяции.

Итак, вероятность того, что расхождение средних сроков службы электроламп в выборке и во всей партии не превысит 8 ч (по абсолютной величине), равна 0,906.

б) Учитывая, что $\gamma = \theta(t,k) = 0{,}95$ и (по таблице) $t_{0,95;16} = 2{,}12$, по **(9.44)** найдем предельную ошибку малой выборки $\Delta_{\text{м.в}} = \dfrac{2{,}12 \cdot 18}{\sqrt{16}} = 9{,}5$ (ч). Теперь по **(9.45)** искомый доверительный интервал $980 - 9{,}5 \le \bar{x}_0 \le 980 + 9{,}5$ или $970{,}5 \le \bar{x}_0 \le 989{,}5$ (ч), т.е. с надежностью 0,95 средний срок службы электроламп в партии заключен от 970,5 до 989,5 ч. ▶

Построение доверительного интервала для генеральной доли по малой выборке. Если доля признака в генеральной совокупности равна p, то вероятность того, что в повторной выборке объема n m элементов обладают этим признаком, определяется по формуле Бернулли: $P_{m,n} = C_n^m p^m q^{n-m}$, где $q = 1-p$, т.е. распределение повторной выборки описывается биномиальным распределением. Так как при $p \neq 0{,}5$ биномиальное распределение несимметрично, то в качестве доверительного интервала для p берут такой интервал (p_1, p_2), что вероятность попадания левее p_1 и правее p_2 одна и та же и равна $(1-\gamma)/2$:

$$\sum_{m=\tilde{m}}^{n} C_n^m p_1^m (1-p_1)^{n-m} = \frac{1-\gamma}{2}; \quad \sum_{m=0}^{\tilde{m}} C_n^m p_2^m (1-p_2)^{n-m} = \frac{1-\gamma}{2},$$

где $\tilde{m} = nw$— фактическое число элементов выборки, обладающих признаком.

Решение таких уравнений можно упростить, если использовать специальные графики, позволяющие при данном объеме выборки n и заданной доверительной вероятности γ определить границы доверительного интервала для генеральной доли p. В качестве примера на рис. 9.3 приведены такие графики для $\gamma = 0{,}9$.

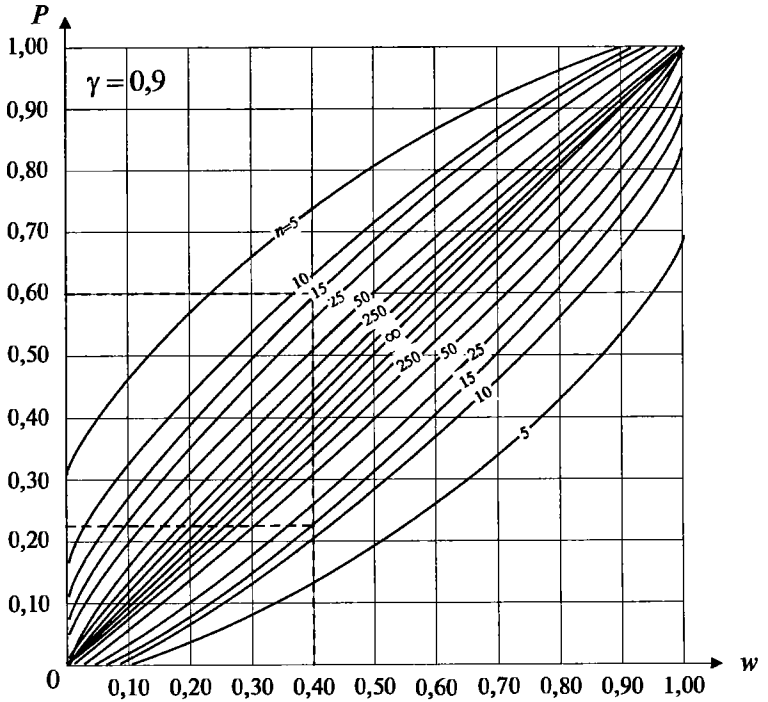


Рис. 9.3

▷ **Пример 9.16.** Опрос случайно отобранных 15 жителей города показал, что 6 из них будут поддерживать действующего мэра на предстоящих выборах. Найти границы, в которых с надежностью 0,9 заключена доля граждан города, которые будут поддерживать на предстоящих выборах действующего мэра.

Р е ш е н и е. Выборочная доля жителей, поддерживающих мэра, $w = m/n = 6/15 = 0{,}4$. По рис 9.3 для $\gamma = 0{,}9$ находим при $w = 0{,}4$ и для $n = 15$ по нижнему графику $p_1 = 0{,}23$, а по верхнему — $p_2 = 0{,}60$, т.е. доля жителей города, поддерживающих мэра, с надежностью 0,9 заключена в границах от 0,23 до 0,60. Очевидно, что более точный ответ на вопрос задачи может быть получен при увеличении объема выборки n. ▶

Построение доверительного интервала для генеральной дисперсии.

Пусть распределение признака (случайной величины) X в генеральной совокупности является нормальным $N(\bar{x}_0;\sigma^2)$. Предположим, что математическое ожидание $M(X)=\bar{x}_0$ (генеральная средняя) известно. Тогда выборочная дисперсия повторной выборки $X_1, X_2,..., X_n$:

$$s_*^2=\frac{\sum_{i=1}^{n}(X_i-\bar{x}_0)^2}{n}$$

(ее не следует путать с выборочной дисперсией

$$s^2=\frac{\sum_{i=1}^{n}(X_i-\bar{x})^2}{n}$$

и «исправленной» выборочной дисперсией $\hat{s}^2=\frac{n}{n-1}s^2$: если s_*^2 характеризует вариацию значений признака относительно генеральной средней $\bar{x}_0$, то s^2 и $\hat{s}^2$ — относительно выборочной средней $\bar{x}$).

Рассмотрим статистику

$$\chi^2=\frac{ns_*^2}{\sigma^2}=\sum_{i=1}^{n}\left(\frac{X_i-\bar{x}_0}{\sigma}\right)^2=\sum_{i=1}^{n}t_i^2$$

Учитывая, что в соответствии с **(9.14)** и **(9.15)** $M(X_i)=\bar{x}_0$, $D(X_i)=\sigma^2$, (i = 1,2,..., n), нетрудно показать, что $M(t)$ = 0 и $\sigma_{t_i}^2=D(t_i)=1$.

В § 4.9 отмечено, что распределение суммы квадратов n независимых случайных величин $\sum_{i=1}^{n}t_i^2$, каждая из которых имеет стандартное нормальное распределение $N(0;1)$, представляет *распределение* χ^2 с $k=n$ *степенями свободы.*

Таким образом, статистика $\chi^2=\frac{ns_*^2}{\sigma^2}$ имеет распределение χ^2 с $k=n$ степенями свободы.

Распределение χ^2 не зависит от неизвестных параметров случайной величины X, а зависит лишь от числа степеней свободы k. Кривые распределения для различного числа степеней свободы показаны на рис. 4.16 (§ 4.9).

Плотность вероятности распределения χ^2 имеет сложный вид, и интегрирование ее является весьма трудоемким процессом. Составлены таблицы для вычисления вероятности того, что случайная величина, имеющая χ^2-распределение с k степенями свободы, превысит некоторое критическое значение $\chi^2_{\alpha;k}$, т.е. $m(\chi^2>\chi^2_{\alpha;k})=\alpha$.

В практике выборочного наблюдения математическое ожидание $\bar{x}_0$, как правило, неизвестно, и приходится иметь дело не с s_*^2, а с s^2 или $\hat{s}^2$. Если $X_1, X_2,...,X_n$ — повторная выборка из нормально распределенной генеральной совокупности, то, как уже сказано выше, случайная величина $\frac{ns^2}{\sigma^2}$ (или $\frac{(n-1)\hat{s}^2}{\sigma^2}$) имеет распределение χ^2 с $k=n-1$ степенями свободы. Поэтому для заданной доверительной вероятности γ можно записать:

$$P\left(\chi_1^2<\frac{ns^2}{\sigma^2}<\chi_2^2\right)=\gamma \qquad \textbf{(9.46)}$$

(графически это площадь под кривой распределения между χ_1^2 и χ_2^2, см. рис. 9.4).

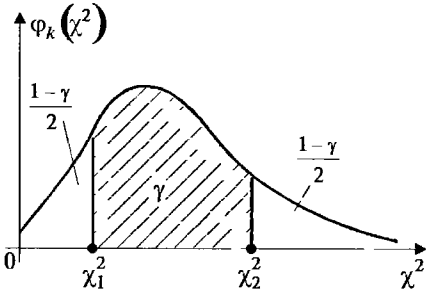


Рис. 9.4

Очевидно, что значения χ_1^2 и χ_2^2 определяются неоднозначно при одном и том же значении заштрихованной площади[1], равной γ. Обычно χ_1^2 и χ_2^2 выбирают таким образом, чтобы вероятности событий $\chi^2 < \chi_1^2$ и $\chi^2 > \chi_2^2$ были одинаковы, т. е.

$$P\left(\chi^2 < \chi_1^2\right) = P\left(\chi^2 > \chi_2^2\right) = \frac{1-\gamma}{2}.$$

Преобразовав двойное неравенство $\chi_1^2 < \frac{ns^2}{\sigma^2} < \chi_2^2$ в равенстве **(9.46)** к равносильному виду $\frac{ns^2}{\chi_2^2} < \sigma^2 < \frac{ns^2}{\chi_1^2}$, получим формулу доверительной вероятности для генеральной дисперсии:

$$P\left(\frac{ns^2}{\chi_2^2} < \sigma^2 < \frac{ns^2}{\chi_1^2}\right) = \gamma, \qquad \textbf{(9.47)}$$

а для среднего квадратического отклонения:

$$P\left(\frac{\sqrt{n}s}{\chi_2} < \sigma < \frac{\sqrt{n}s}{\chi_1}\right) = \gamma. \qquad \textbf{(9.48)}$$

При использовании таблиц значений $\chi_{\alpha;k}^2$, полученных из равенства $P\left(\chi^2 > \chi_{\alpha;k}^2\right) = \alpha$, необходимо учесть, что $P\left(\chi^2 < \chi_1^2\right) = 1 - P\left(\chi^2 > \chi_1^2\right)$, поэтому условие $P\left(\chi^2 < \chi_1^2\right) = \frac{1-\gamma}{2}$ равносильно условию $P\left(\chi^2 > \chi_1^2\right) = 1 - \frac{1-\gamma}{2} = \frac{1+\gamma}{2}$. Таким образом, значения χ_1^2 и χ_2^2 находим по табл. V приложений из равенств:

$$P\left(\chi^2 > \chi_1^2\right) = \frac{1+\gamma}{2}, \qquad \textbf{(9.49)}$$

[1] При построении доверительных интервалов для $\bar{x}_0$ и p мы эту неоднозначность обходили тем, что брали доверительный интервал, симметричный относительно несмещенной точечной оценки. Здесь это смысла не имеет, так как в отличие от выборочных распределений $\bar{x}$ и w распределение χ^2 не обладает симметрией.

$$P\left(\chi^2 > \chi_2^2\right) = \frac{1-\gamma}{2}, \qquad \textbf{(9.50)}$$

т.е. при $k = n - 1$ $\chi_1^2 = \chi_{(1+\gamma)/2;n-1}^2$, $\chi_2^2 = \chi_{(1-\gamma)/2;n-1}^2$.

▷ **Пример 9.17.** На основании выборочных наблюдений производительности труда 20 работниц было установлено, что среднее квадратическое отклонение суточной выработки составляет 15 м ткани в час. Предполагая, что производительность труда работницы имеет нормальное распределение, найти границы, в которых с надежностью 0,9 заключены генеральные дисперсия и среднее квадратическое отклонение суточной выработки работниц.

Р е ш е н и е. Имеем $\gamma = 0{,}9; (1-\gamma)/2 = 0{,}05$. $(1+\gamma)/2 = 0{,}95$.

При числе степеней свободы $k=n-1=20-1=19$ в соответствии с **(9.49)** и **(9.50)** определим χ_1^2 и χ_2^2 по табл. V приложений: $\chi_1^2 = \chi_{0,95;19}^2 = 10{,}1$ и $\chi_2^2 = \chi_{0,05;19}^2 = 30{,}1$. Тогда доверительный интервал для σ^2 по **(9.47)** можно записать в виде: $\frac{20}{30{,}1} \cdot 15^2 < \sigma^2 < \frac{20}{10{,}1} \cdot 15^2$ или $149{,}5 < \sigma^2 < 445{,}6$ и для σ по **(9.48)**: $\sqrt{149{,}5} < \sigma < \sqrt{445{,}6}$ или $12{,}2 < \sigma < 21{,}1$ (м/ч).

Итак, с надежностью 0,9 дисперсия суточной выработки работниц заключена в границах от 149,5 до 445,6, а ее среднее квадратическое отклонение — от 12,2 до 21,1 метров ткани в час. ▶

З а м е ч а н и е. Таблица значений $\chi_{\alpha;k}^2$ (прил. V) составлена при числе степеней свободы k от 1 до 30. При $k>30$ можно считать (см. § 4.9), что случайная величина $\sqrt{2\chi^2} - \sqrt{2k-1}$ имеет стандартное нормальное распределение $N(0;1)$. Поэтому для определения χ_1^2 и χ_2^2 следует записать, что

$$P\left(\left|\sqrt{2\chi^2} - \sqrt{2k-1}\right| < t\right) = \Phi(t) = \gamma, \qquad \textbf{(9.51)}$$

откуда $-t < \sqrt{2\chi^2} - \sqrt{2k-1} < t$ и после преобразований $\frac{1}{2}\left(\sqrt{2k-1} - t\right)^2 < \chi^2 < \frac{1}{2}\left(\sqrt{2k-1} + t\right)^2$. Таким образом, при расчете доверительного

интервала при $k > 30$ надо полагать $\chi_1^2 = \frac{1}{2}\left(\sqrt{2k-1} - t\right)^2$, $\chi_2^2 = \frac{1}{2}\left(\sqrt{2k-1} + t\right)^2$, где $\Phi(t) = \gamma$.

▷ **Пример 9.18.** Решить задачу, приведенную в примере **9.17**, при $n = 100$ работницам.

Р е ш е н и е. При $\gamma = \Phi(t) = 0{,}9$ по таблице II приложений $t = 1{,}645$, поэтому

$\chi_1^2 = \frac{1}{2}\left(\sqrt{2 \cdot 99 - 1} - 1{,}645\right)^2 = 76{,}8$, $\chi_2^2 = \frac{1}{2}\left(\sqrt{2 \cdot 99 - 1} + 1{,}645\right)^2 = 122{,}9$.

Далее решение, аналогичное примеру **9.17**, приводит к доверительным интервалам для σ^2: $183{,}1 < \sigma^2 < 293{,}0$ и для σ: $13{,}5 < \sigma < 17{,}1$ (м/ч).

Упражнения

9.19. Для исследования доходов населения города, составляющего 20 тыс. человек, по схеме собственно-случайной бесповторной выборки было отобрано 1000 жителей. Получено следующее распределение жителей по месячному доходу (руб.):

x_i	менее 500	500—1000	1000—1500	1500—2000	2000—2500	свыше 2500
n_i	58	96	239	328	147	132

Необходимо: 1. а) Найти вероятность того, что средний месячный доход жителя города отличается от среднего дохода его в выборке не более, чем на 45 руб. (по абсолютной величине); б) определить границы, в которых с надежностью 0,99 заключен средний месячный доход жителей города. 2. Каким должен быть объем выборки, чтобы те же границы гарантировать с надежностью 0,9973?

9.20. Решить пример **9.19** при условии, что население города неизвестно, а известно лишь, что оно очень большое по сравнению с объемом выборки.

9.21. По данным примера **9.19** необходимо: 1. а) Найти вероятность того, что доля малообеспеченных жителей города (с доходом менее 500 руб.) отличается от доли таких же жителей в выборке не более, чем на 0,01 (по абсолютной величине); б) определить границы, в которых с надежностью 0,98 заключена доля малообеспеченных жителей города. 2. Каким должен быть объем выборки, чтобы те же границы для доли малообеспеченных жителей города гарантировать с надежностью 0,9973? 3. Как изменились бы результаты, полученные в п. 1. а) и 2, если бы о доле малообеспеченных жителей вообще не было ничего известно?

9.22. Решить пример **9.21** при условии, что население города неизвестно, а известно лишь, что оно очень большое по сравнению с объемом выборки.

9.23. Из 5000 вкладчиков банка по схеме случайной бесповторной выборки было отобрано 300 вкладчиков. Средний размер вклада в выборке составил 8000 руб., а среднее квадратическое отклонение 2500 руб. Какова вероятность того, что средний размер вклада случайно выбранного вкладчика отличается от его среднего размера в выборке не более, чем на 100 руб. (по абсолютной величине)?

9.24. В результате выборочного наблюдения получены следующие данные о часовой выработке (в ед./ч) 50 рабочих, отобранных из 1000 рабочих цеха:

Часовая выработка	0,9	1,1	1,3	1,5	1,7	1,9
Число рабочих	1	2	10	17	16	4

1) Найти (с надежностью 0,95) максимальное отклонение средней часовой выработки рабочих в выборке от средней во всем цехе (по абсолютной величине), если выборка: а) повторная; б) бесповторная. 2) Найти объем выборки, при котором с надежностью 0,99 можно гарантировать вдвое меньшее максимальное отклонение тех же характеристик.

9.25. Из партии, содержащей 8000 телевизоров, отобрано 800. Среди них оказалось 10% не удовлетворяющих стандарту. Найти границы, в которых с вероятностью 0,95 заключена доля телевизоров, удовлетворяющих стандарту, во всей партии для повторной и бесповторной выборок.

9.26. По результатам социологического обследования при опросе 1500 респондентов рейтинг президента (т.е. процент опрошенных, одобряющих его деятельность) составил 30%. Найти границы, в которых с надежностью 0,95 заключен рейтинг президента (при опросе

всех жителей страны). Сколько респондентов надо опросить, чтобы с надежностью 0,99 гарантировать предельную ошибку социологического обследования не более 1%? Тот же вопрос, если никаких данных о рейтинге президента нет.

9.27. Каким должен быть объем выборки, отобранной по схеме случайной бесповторной выборки из партии, содержащей 8000 деталей, чтобы с вероятностью 0,994 можно было утверждать, что доли первосортных деталей в выборке и во всей партии отличаются не более чем на 0,05 (по абсолютной величине)? Задачу решить для случаев: а) о доле первосортных деталей во всей партии ничего не известно; б) их не более 80%.

9.28. Производятся независимые испытания с одинаковой, но неизвестной вероятностью p появления события A в каждом испытании. Найти доверительный интервал для оценки вероятности p с надежностью $\gamma = 0{,}95$, если в $n = 60$ испытаниях событие A появилось $m = 15$ раз.

9.29. Решить пример **9.28** при $\gamma = 0{,}9$; $n = 10$; $m = 2$.

9.30. Из большой партии по схеме случайной повторной выборки было проверено 150 изделий с целью определения процента влажности древесины, из которой изготовлены эти изделия. Получены следующие результаты:

Процент влажности	11—13	13—15	15—17	17—19	19—21
Число изделий	8	42	51	37	12

Считая, что процент влажности изделия — случайная величина, распределенная по нормальному закону, найти: а) вероятность того, что средний процент влажности заключен в границах от 12,5 до 17,5; б) границы, в которых с вероятностью 0,95 будет заключен средний процент влажности изделий во всей партии.

9.31. По данным 9 измерений некоторой величины найдены средняя результатов измерений $x = 30$ и выборочная дисперсия $s^2 = 36$. Найти границы, в которых с надежностью 0,99 заключено истинное значение измеряемой величины.

9.32. Произведено 12 измерений одним прибором (без систематической ошибки) некоторой величины, имеющей нормальное распределение, причем выборочная дисперсия случайных ошибок измерений оказалась равной 0,36. Найти границы, в которых с надежностью 0,95 заключено среднее квадратическое отклонение случайных ошибок измерений, характеризующих точность прибора.

9.33. Решить пример **9.32** при $n = 100$ измерениях.

9.34. Распределение 200 элементов (устройств) по времени безотказной работы (в часах) представлено в таблице:

Время безотказной работы	0—5	5—10	10—15	15—20	20—25	25—30
Число устройств	133	45	15	4	2	1

Предполагая, что время безотказной работы элементов имеет показательный закон распределения, найти: а) вероятность того, что время безотказной работы будет заключено в пределах от 3 до 8 ч; б) границы, в которых с надежностью 0,95 будет заключено среднее время безотказной работы элементов.

Указание. В качестве оценки параметра λ взять величину, обратную выборочной средней.

Глава 10 Проверка статистических гипотез

10.1. Принцип практической уверенности

Прежде чем перейти к рассмотрению понятия статистической гипотезы, сформулируем так называемый **принцип практической уверенности**, лежащий в основе применения выводов и рекомендаций с помощью теории вероятностей и математической статистики:

Если вероятность события A в данном испытании очень мала, то при однократном выполнении испытания можно быть уверенным в том, что событие A не произойдет, и в практической деятельности вести себя так, как будто событие A вообще невозможно.

Этот принцип не может быть доказан математически; он подтверждается всем практическим опытом человеческой деятельности, и мы постоянно (хотя и бессознательно) им руководствуемся. Например, отправляясь самолетом в другой город, мы не рассчитываем на возможность погибнуть в авиационной катастрофе, хотя некоторая (весьма малая) вероятность такого события все же имеется.

Обратим внимание на то, что принцип практической уверенности о невозможности маловероятных событий сформулирован «при однократном выполнении испытания». Если же произведено много испытаний, в каждом из которых вероятность события A даже очень мала, то существенно повышается вероятность того, что событие A произойдет хотя бы один раз в массе испытаний. Действительно, пусть вероятность $P(A) = \alpha$, где $\alpha << 1$. Тогда вероятность события B, состоящего в том, что событие A произойдет хотя бы один раз в n независимых испытаниях, по формуле (**1.29**) равна (при $\alpha << 1$):

$$P(B) = 1 - (1-\alpha)^n \approx 1 - (1 - n\alpha) = n\alpha,$$

т.е. вероятность $P(B)$ увеличилась по сравнению с $P(A)$ в n раз.

Таким образом, при многократном повторении испытаний мы уже не можем считать маловероятное событие A практически невозможным.

Вопрос о том, насколько мала должна быть вероятность α события A, чтобы его можно было считать практически невозможным, выходит за рамки математической теории и решается в каждом отдельном случае с учетом важности последствий, вытекающих из наступления события A. В одних случаях считается возможным пренебрегать событиями, имеющими вероятность меньше 0,05, а в других, когда речь идет, например, о разрушении сооружений, гибели судна и т.п., нельзя пренебрегать событиями, которые могут появиться с вероятностью, равной 0,001.

10.2. Статистическая гипотеза и общая схема ее проверки

С теорией статистического оценивания параметров тесно связана проверка статистических гипотез. Она используется всякий раз, когда необходим обоснованный вывод о преимуществах того или иного способа инвестиций, измерений, стрельбы, технологического процесса, об эффективности нового метода обучения, управления, о пользе вносимого удобрения, лекарства, об уровне доходности ценных бумаг, о значимости математической модели и т.д.

Определение. ***Статистической гипотезой*** *называется любое предположение о виде или параметрах неизвестного закона распределения.*

Различают *простую* и *сложную* статистические гипотезы. Простая гипотеза, в отличие от сложной, полностью определяет теоретическую функцию распределения случайной величины. Например, гипотезы «вероятность появления события в схеме Бернулли равна 1/2», «закон распределения случайной величины нормальный с параметрами $a = 0$, $\sigma^2 = 1$» являются простыми, а гипотезы «вероятность появления события в схеме Бернулли заключена между 0,3 и 0,6», «закон распределения не является нормальным» — сложными.

Проверяемую гипотезу обычно называют *нулевой* (или *основной*) и обозначают H_0. Наряду с нулевой гипотезой H_0 рассматривают *альтернативную*, или *конкурирующую*, гипотезу H_1, являющуюся логическим отрицанием H_0. *Нулевая и альтернативная гипотезы представляют собой две возможности выбора, осуществляемого в задачах проверки статистических гипотез*

Суть проверки статистической гипотезы заключается в том, что используется специально составленная выборочная характеристика (*статистика*) $\tilde{\theta}_n(x_1,...,x_n)$, полученная по выборке $X_1,...,X_n$, точное или приближенное распределение которой известно. Затем по этому выборочному распределению определяет-

ся критическое значение $\theta_{кр}$ — такое, что если гипотеза H_0 верна, то вероятность $P(\tilde{\theta}_n > \theta_{кр}) = \alpha$ мала; так что в соответствии с принципом практической уверенности в условиях данного исследования событие $\tilde{\theta}_n > \theta_{кр}$ можно (с некоторым риском) считать практически невозможным. Поэтому, если в данном конкретном случае обнаруживается отклонение $\tilde{\theta}_n > \theta_{кр}$, то гипотеза H_0 отвергается, в то время как появление значения $\tilde{\theta}_n \le \theta_{кр}$ считается совместимым с гипотезой H_0, которая тогда принимается (точнее, не отвергается). *Правило, по которому гипотеза H_0 отвергается или принимается, называется* ***статистическим критерием*** или ***статистическим тестом.***

Таким образом, множество возможных значений статистики критерия (критической статистики) $\tilde{\theta}_n$ разбивается на два непересекающихся подмножества: *критическую область* (*область отклонения гипотезы*) W и *область допустимых значений* (*область принятия гипотезы*) $\overline{W}$. Если фактически наблюдаемое значение статистики критерия $\tilde{\theta}_n$ попадает в критическую область W, то гипотезу H_0 отвергают. При этом возможны четыре случая (табл. 10.1).

Таблица 10.1

Гипотеза H_0	*Принимается*	*Отвергается*
Верна	Правильное решение	Ошибка 1-го рода
Неверна	Ошибка 2-го рода	Правильное решение

О п р е д е л е н и е. *Вероятность α допустить ошибку 1-го рода, т.е. отвергнуть гипотезу H_0, когда она верна, называется* ***уровнем значимости***, или ***размером, критерия***[1].

Вероятность допустить ошибку 2-го рода, т.е. принять гипотезу H_0, когда она неверна, обычно обозначают β.

О п р е д е л е н и е. *Вероятность $(1-\beta)$ не допустить ошибку 2-го рода, т.е. отвергнуть гипотезу H_0, когда она неверна, называется* ***мощностью*** (*или* ***функцией мощности***) ***критерия.***

1 Вероятность $1-\alpha$ не допустить ошибку первого рода, т.е. принять гипотезу H_0, когда она верна, иногда называют *оперативной характеристикой критерия*

Пользуясь терминологией статистического контроля качества продукции, можно сказать, что вероятность α представляет «риск поставщика», связанный с забраковкой по результатам выборочного контроля изделий всей партии, удовлетворяющей стандарту, а вероятность β — «риск потребителя», связанный с принятием по анализу выборки партии, не удовлетворяющей стандарту.

Применяя юридическую терминологию, α — вероятность вынесения судом обвинительного приговора, когда на самом деле обвиняемый невиновен, β — вероятность вынесения судом оправдательного приговора, когда на самом деле обвиняемый виновен в совершении преступления. В ряде прикладных исследований ошибка первого рода α означает вероятность того, что предназначавшийся наблюдателю сигнал не будет им принят, а ошибка второго рода β — вероятность того, что наблюдатель примет ложный сигнал.

Возможностью двойной ошибки (1-го и 2-го рода) проверка гипотез отличается от рассматриваемого выше интервального оценивания параметров, в котором имелась лишь одна возможность ошибки: получение доверительного интервала, который на самом деле не содержит оцениваемого параметра.

Вероятности ошибок 1-го и 2-го рода (α и β) *однозначно определяются выбором критической области.* Очевидно, желательно сделать как угодно малыми α и β. Однако это противоречивые требования: при фиксированном объеме выборки можно сделать как угодно малой лишь одну из величин — α или β, что сопряжено с неизбежным увеличением другой. Лишь при увеличении объема выборки возможно о д н о в р е м е н н о е уменьшение вероятностей α и β (см. пример **10.0**).

Какими принципами следует руководствоваться при построении критической области W?

Предположим, что используемая для проверки (тестирования) нулевой гипотезы H_0 статистика критерия $\tilde{\theta}_n$ имеет нормальный закон распределения $N(a_0; \sigma^2)$. В качестве критической области, отвечающей уровню значимости α=0,05, можно взять множество областей — таких, что площадь соответствующих им криволинейных трапеций под кривой распределения составляет 5/100 от общей площади под кривой распределения. Например (рис. 10.1): [I] — область больших положительных отклонений (при $\tilde{\theta}_n > \theta_{кр\,1}$); [II] — область больших отрицательных отклоне-

ний (при $\tilde{\theta}_n < \theta_{кр.2}$); [III] — область больших по абсолютной величине отклонений (при $\tilde{\theta}_n < \theta'_{кр.3}$, $\tilde{\theta}_n > \tilde{\theta}''_{кр\,3}$); [IV] — область малых по абсолютной величине отклонений (при $\theta_{кр.4} < \tilde{\theta}_n < \theta''_{кр.4}$) и т.д.

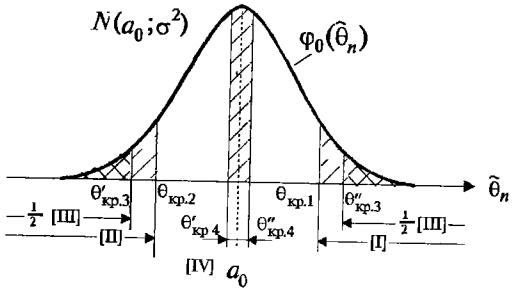


Рис. 10.1

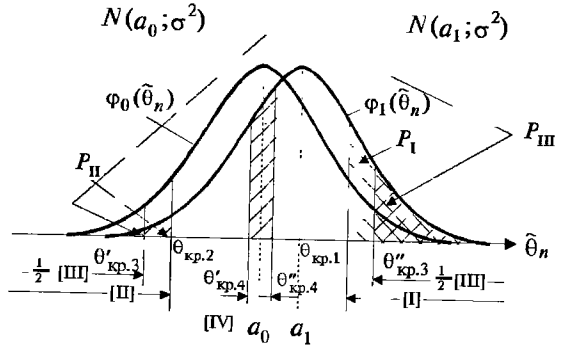


Рис. 10.2

Какую из этих областей предпочесть в качестве критической? Пусть с проверяемой гипотезой H_0 конкурирует другая, альтернативная, гипотеза H_1, при которой распределение статистики критерия $\tilde{\theta}_n$ нормально: $N(a_0;\ \sigma^2)$, где $a_1 > a_0$ (рис. 10.2). Очевидно, *следует предпочесть ту критическую область, при которой мощность критерия будет наибольшей.* Если, например, критическая область типа [I], то в случае $\tilde{\theta}_n < \theta_{кр.1}$ гипотеза H_0 принимается. Но в этом случае может быть верна конкурирующая гипотеза H_1 с вероятностью ошибки второго рода β. Вероятность β интерпретируется площадью под кривой распределения $\varphi_1(\tilde{\theta}_n)$ левее $\theta_{кр.1}$[1], а мощность критерия $(1-\beta)$ — площадью P_I правее $\theta_{кр.1}$ (см. рис. 10.2). Аналогично P_{II}, P_{III}, P_{IV} интерпретируют мощность критерия при критических областях соответственно II, III и IV типов (на рис. 10.2 площади P_I—P_{IV} заштрихованы)[2]. Очевидно, что в данном случае целесообразно выбрать в качестве критической область [I], т.е. правостороннюю критическую область, так как такой выбор гарантирует максимальную мощность критерия.

Требования к критической области аналитически можно записать так:

$$P\left(\tilde{\theta}_n \in W/H_0\right) = \alpha,$$
$$P\left(\tilde{\theta}_n \in W/H_1\right) = \max, \qquad \textbf{(10.1)}$$

т.е. *критическую область W следует выбирать так, чтобы вероятность попадания в нее статистики критерия $\tilde{\theta}_n$ была минимальной и равной α, если верна нулевая гипотеза H_0, и максимальной в противоположном случае.*

Другими словами, *критическая область должна быть такой, чтобы при заданном уровне значимости α мощность критерия $1-\beta$ была максимальной.* Задача построения такой критической области W (или, как говорят, построения *наиболее мощного критерия*) для простых гипотез решается с помощью следующей теоремы.

Теорема (лемма) Неймона—Пирсона. *Среди всех критериев заданного уровня значимости α, проверяющих простую гипотезу H_0 против альтернативной гипотезы H_1, критерий отношения правдоподобия является наиболее мощным.*

Поясним смысл этой теоремы, полагая случайную величину X непрерывной.

Если верна простая гипотеза H_0, то плотность вероятности $\varphi(x)$ определяется однозначно, и функция правдоподобия $L_0(x)$, выражающая плотность вероятности совместного появления результатов выборки $(x_1, x_2, ..., x_n)$, имеет вид (см. § 9.3):

$$L_0(x_1, ...,\ x_n) = \varphi_0(x_1)\varphi_0(x_2)...\varphi_0(x_n).$$

---

[1] Здесь отчетливо видно, что если увеличить $\theta_{кр.1}$, то ошибка α 1-го рода уменьшится (станет меньше, чем 0,05), но увеличится ошибка 2-го рода β, и наоборот; одновременно же уменьшить и α, и β невозможно.

[2] P_{III} частично перекрывается с P_I и P_{II}.

Напомним, что функция $L_0(x_1, ..., x_n)$ есть *мера правдоподобности* получения выборочных наблюдений $x_1, x_2, ..., x_n$.

Аналогично, если верна простая гипотеза H_1, то функция правдоподобия

$$L_1(x_1, ..., x_n) = \varphi_1(x_1)\varphi_1(x_2) ... \varphi_1(x_n).$$

В теореме Неймана—Пирсона рассматривается отношение правдоподобия L_1/L_0 (при $L_0 \neq 0$); чем правдоподобнее выборка в условиях гипотезы H_1, тем больше отношение L_1/L_0 или его логарифм $\ln(L_1/L_0)$. А критерий этого отношения, по заключению теоремы, и является наиболее мощным среди других возможных критериев.

Используя данный критерий, можно найти такую постоянную C (или $\ln C = c$), что

$$P\left(\frac{L_1(x_1,...,x_n)}{L_0(x_1,..., x_n)} > C\right) = P\left(\ln\frac{L_1(x_1,...,x_n)}{L_0(x_1,..., x_n)} > c\right) = \alpha.$$

С помощью полученной постоянной C (или c) определяется критическая область W критерия и его мощность.

Пример 10.0. Случайная величина X имеет нормальный закон распределения $N(a; \sigma^2)$, где $a = M(X)$ не известно, а $\sigma^2 = D(X)$ известно. Построить наиболее мощный критерий проверки гипотезы H_0: $a = a_0$ против альтернативной H_1: $a = a_1 > a_0$. Найти: а) мощность критерия; б) минимальный объем выборки, обеспечивающий заданные уровень значимости α и мощность критерия $1 - \beta$.

Р е ш е н и е. Если верна гипотеза H_0, т.е. $X \sim N(a_0; \sigma^2)$, то функция правдоподобия (см. § 9.3) имеет вид:

$$L_0(x_1,...,x_n) = \frac{1}{\sigma^n(2\pi)^{n/2}} e^{-\frac{\sum_{i=1}^{n}(x_i-a_0)^2}{2\sigma^2}}.$$

Аналогично, если верна гипотеза H_1, т.е. $X \sim N(a_1; \sigma^2)$, то

$$L_1(x_1,...,x_n) = \frac{1}{\sigma^n(2\pi)^{n/2}} e^{-\frac{\sum_{i=1}^{n}(x_i-a_1)^2}{2\sigma^2}}.$$

Согласно теореме Неймана—Пирсона наиболее мощный критерий основан на отношении правдоподобия L_1/L_0. Найдем его логарифм; получим

$$\ln(L_1/L_0) = -\frac{\sum_{i=1}^{n}(x_i-a_1)^2}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i-a_0)^2}{2\sigma^2} =$$

$$= \frac{1}{2\sigma^2}\sum_{i=1}^{n}[2x_i(a_1-a_0) - (a_1^2 - a_0^2)] = \frac{1}{2\sigma^2}(a_1-a_0)\sum_{i=1}^{n}(2x_i - a_1 - a_0) =$$

$$= \frac{1}{2\sigma^2}(a_1-a_0)(2\bar{x} - a_1 - a_0)n, \text{ ибо } \bar{x} = \sum_{i=1}^{n} x_i/n.$$

Для построения критерия найдем такую постоянную C (или $\ln C = c$), что

$$P\left(\frac{L_1}{L_0} > C\right) = P\left(\ln\frac{L_1}{L_0} > c\right) = \alpha.$$

Полученное выражение для уровня значимости α можно заменить ему равносильным (учитывая монотонность функции $\ln(L_1/L_0)$ относительно $\bar{x}$):

$$P(\bar{x} > c') = \alpha.$$

Для определения c' следует учесть, что если случайная величина X распределена нормально, т.е. $X \sim N(a_0, \sigma^2)$, то ее средняя $\bar{x}$ также распределена нормально с параметрами a_0 и σ^2/n (см. § 6.3, 9.3), т.е. $\bar{x} \sim N(a_0, \sigma^2/\sqrt{n})$.

Используя выражение функции распределения нормального закона через функцию Лапласа (**4.30**), получим

$$P(\bar{x} > c') = 1 - P(\bar{x} \le c') = 1 - \left[\frac{1}{2} + \frac{1}{2}\Phi\left(\frac{c'-a_0}{\sigma}\sqrt{n}\right)\right] =$$

$$= \frac{1}{2} - \frac{1}{2}\Phi\left(\frac{c'-a_0}{\sigma}\sqrt{n}\right) = \alpha,$$

откуда $\Phi\left(\frac{c'-a_0}{\sigma}\sqrt{n}\right) = 1 - 2\alpha$ или $\frac{c'-a_0}{\sigma}\sqrt{n} = t_{1-2\alpha}$ и определяющее границу критической области W значение $c' = a_0 + t_{1-2\alpha}\frac{\sigma}{\sqrt{n}}$.

Следовательно, наиболее мощным критерием проверки гипотезы H_0: $a = a_0$ против альтернативной H_1: $a = a_1 > a_0$ является следующий: *гипотеза* H_0 *отвергается, если* $\bar{x} > a_0 + t_{1-2\alpha}\frac{\sigma}{\sqrt{n}}$;

H_0 *не отвергается, если* $\bar{x} \le a_0 + t_{1-2\alpha}\frac{\sigma}{\sqrt{n}}$.

а) Для нахождения мощности критерия определим вначале вероятность β допустить ошибку 2-го рода — принять гипотезу

H_0, когда она не верна, а верна альтернативная гипотеза H_1, т.е. $X \sim N(a_1, \sigma^2)$ или $\bar{x} \sim N(a_1, \sigma^2/\sqrt{n})$:

$$\beta = P\left(\bar{x} \le a_0 + t_{1-2\alpha}\frac{\sigma}{\sqrt{n}}\right) = \frac{1}{2} + \frac{1}{2}\Phi\left(\frac{a_0 + t_{1-2\alpha}\sigma/\sqrt{n} - a_1}{\sigma}\sqrt{n}\right) =$$

$$= \frac{1}{2} - \frac{1}{2}\Phi\left(\frac{(a_1 - a_0)\sqrt{n}}{\sigma} - t_{1-2\alpha}\right).$$

Следовательно, мощность критерия есть

$$1 - \beta = \frac{1}{2} + \frac{1}{2}\Phi\left(\frac{(a_1 - a_0)\sqrt{n}}{\sigma} - t_{1-2\alpha}\right).$$

Рассматривая полученные выражения, еще раз (теперь уже аналитически) убеждаемся в том, что уменьшение уровня значимости α при неизменном объеме выборки n ведет к увеличению вероятности β и соответственно к снижению мощности критерия $1 - \beta$. И только при увеличении объема выборки n возможно, уменьшая вероятность α, одновременно уменьшать вероятность β (увеличивать мощность критерия $1 - \beta$).

б) При заданных вероятностях ошибок 1-го и 2-го рода α и β из выражения для β нетрудно найти соответствующий объем выборки по формуле:

$$n = \frac{(t_{1-2\alpha} + t_{1-2\beta})^2 \sigma^2}{(a_1 - a_0)^2} \quad \blacktriangleright \qquad \textbf{(10.1')}$$

В зависимости от вида конкурирующей гипотезы H_1 выбирают *правостороннюю, левостороннюю* или *двустороннюю* критическую область. Так, в рассмотренном примере мы убедились, что при конкурирующей гипотезе H_1: $a_1 > a_0$ следовало использовать правостороннюю критическую область [I] (см. рис. 10.1, 10.2). Аналогично можно показать, что в случае H_1: $a_1 < a_0$ следовало использовать левостороннюю критическую область [II], а при гипотезе H_1: $a_1 \ne a_0$ — двустороннюю критическую область [III]. Границы критических областей $\theta_{кр}$ при заданном уровне значимости α определяются соответственно из соотношений:

для правосторонней критической области

$$P\left(\tilde{\theta}_n > \theta_{кр}\right) = \alpha, \qquad \textbf{(10.2)}$$

для левосторонней критической области

$$P\left(\tilde{\theta}_n < \theta_{кр}\right) = \alpha, \qquad \textbf{(10.3)}$$

для двусторонней критической области

$$P\left(\tilde{\theta}_n < \theta_{кр.1}\right) = P\left(\tilde{\theta}_n > \theta_{кр.2}\right) = \frac{\alpha}{2}. \qquad \textbf{(10.4)}$$

Следует отметить, что в компьютерных статистических пакетах обычно не находятся границы критической области $\theta_{кр}$, необходимые для сравнения их с фактически наблюдаемыми значениями выборочных характеристик $\tilde{\theta}_{набл}$ и принятия решения о справедливости гипотезы H_0. А рассчитывается точное значение уровня значимости (*p-valio*) исходя из соотношения $P(\tilde{\theta}_n > \tilde{\theta}_{набл.}) = p$. Если p очень мало, то гипотезу H_0 отвергают, в противном случае H_0 принимают (точнее, не отвергают; при этом рассчитанное на компьютере значение p может быть удвоено при выборе двусторонней критической области).

Принцип проверки статистической гипотезы не дает логического доказательства ее верности или неверности. Принятие гипотезы H_0 в сравнении с альтернативной H_1 не означает, что мы уверены в абсолютной правильности H_0 или что высказанное в гипотезе H_0 утверждение является наилучшим, единственно подходящим; просто гипотеза H_0 не противоречит имеющимся у нас выборочным данным, таким же свойством наряду с H_0 могут обладать и другие гипотезы. Более того, возможно, что при увеличении объема выборки n либо при испытании H_0 против другой альтернативной гипотезы H_2 гипотеза H_0 будет отвергнута. Так что *принятие гипотезы H_0 следует расценивать не как раз и навсегда установленный, абсолютно верный содержащийся в ней факт, а лишь как достаточно правдоподобное, не противоречащее опыту утверждение.*

В описанной выше схеме проверка гипотез основывается на предположении об известном законе распределения генеральной совокупности, из которого следует определенное распределение критерия. Критерии проверки таких гипотез называются *параметрическими.* Если закон распределения генеральнои совокупности неизвестен, то соответствующие критерии получили название *непараметрических.* Естественно, что непараметрические критерии обладают значительно меньшей мощностью, чем параметрические. Это означает, что для сохранения той же мощности при использовании непараметрического критерия по сравнению с параметрическим нужно иметь значительно больший объем наблюдений.

По своему прикладному содержанию статистические гипотезы можно подразделить на несколько основных типов:

- о равенстве числовых характеристик генеральных совокупностей;
- о числовых значениях параметров;
- о законе распределения;
- об однородности выборок (т.е. принадлежности их одной и той же генеральной совокупности);
- о стохастической независимости элементов выборки.

10.3. Проверка гипотез о равенстве средних двух и более совокупностей

Сравнение средних двух совокупностей имеет важное практическое значение. На практике часто встречается случай, когда средний результат одной серии экспериментов отличается от среднего результата другой серии. При этом возникает вопрос, можно ли объяснять обнаруженное расхождение средних неизбежными случайными ошибками эксперимента или оно вызвано некоторыми закономерностями[1]. В промышленности задача сравнения средних часто возникает при выборочном контроле качества изделий, изготовленных на разных установках или при различных технологических режимах, в финансовом анализе — при сопоставлении уровня доходности различных активов и т.д.

Сформулируем задачу. Пусть имеются две совокупности, характеризуемые генеральными средними $\bar{x}_0$ и $\bar{y}_0$ и **известными дисперсиями** σ_x^2 и σ_y^2. Необходимо проверить гипотезу H_0 о равенстве генеральных средних, т.е. H_0: $\bar{x}_0 = \bar{y}_0$. Для проверки гипотезы H_0 из этих совокупностей взяты две независимые выборки объемов n_1 и n_2, по которым найдены средние арифметические $\bar{x}$ и $\bar{y}$ и выборочные дисперсии s_x^2 и s_y^2.

При достаточно больших объемах выборки, как отмечено в § 9.6, выборочные средние $\bar{x}$ и $\bar{y}$ имеют приближенно нормальный закон распределения, соответственно $N\left(\bar{x}_0, \sigma_x^2\right)$ и $N\left(\bar{y}_0, \sigma_y^2\right)$.

В случае справедливости гипотезы H_0 разность $\bar{x} - \bar{y}$ имеет нормальный закон распределения с математическим ожиданием

[1] Поэтому проверку гипотез такого типа называют *проверкой (оценкой) значимости (существенности) различия выборочных средних* или других характеристик.

$M(\bar{x}-\bar{y}) = M(\bar{x}) - M(\bar{y}) = \bar{x}_0 - \bar{y}_0 = 0$ и дисперсией $\sigma^2_{\bar{x}-\bar{y}} = \sigma^2_{\bar{x}} + \sigma^2_{\bar{y}} = \dfrac{\sigma_x^2}{n_1} + \dfrac{\sigma_y^2}{n_2}$ (напомним, что дисперсия разности независимых случайных величин равна сумме их дисперсий, а дисперсия средней n независимых слагаемых в n раз меньше дисперсии каждого).

Поэтому при выполнении гипотезы H_0 статистика

$$t = \frac{(\bar{x}-\bar{y}) - M(\bar{x}-\bar{y})}{\sigma_{\bar{x}-\bar{y}}} = \frac{\bar{x}-\bar{y}}{\sqrt{\dfrac{\sigma_x^2}{n_1} + \dfrac{\sigma_y^2}{n_2}}} \qquad \textbf{(10.5)}$$

имеет стандартное нормальное распределение $N(0;1)$.

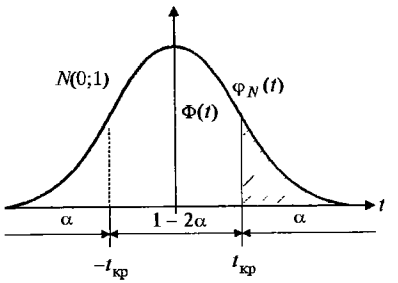


Рис. 10.3

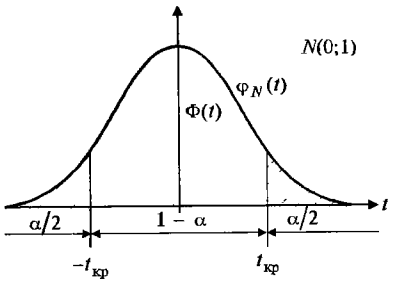


Рис. 10.4

Согласно **(10.2)**—**(10.4)** в случае конкурирующей гипотезы H_1: $\bar{x}_0 > \bar{y}_0$ (или H_1: $\bar{x}_0 < \bar{y}_0$) выбирают *одностороннюю* критическую область и критическое значение статистики находят из условия (рис. **10.3**)

$$\Phi\left(t_{кр}\right) = \Phi\left(t_{1-2\alpha}\right) = 1 - 2\alpha\,, \qquad \textbf{(10.6)}$$

а при конкурирующей гипотезе H_2: $\bar{x}_0 \neq \bar{y}_0$ выбирают *двустороннюю* критическую область и критическое значение статистики находят из условия (рис. **10.4**)

$$\Phi\left(t_{кр}\right) = \Phi\left(t_{1-\alpha}\right) = 1 - \alpha\,. \qquad \textbf{(10.7)}$$

Если фактически наблюдаемое значение статистики t больше критического $t_{кр}$, определенного на уровне значимости α (по абсолютной величине), т.е. $|t| > t_{кр}$, то гипотеза H_0 отвергается. Если $|t| \le t_{кр}$, то делается вывод, что нулевая гипотеза H_0 не противоречит имеющимся наблюдениям.

▷ **Пример 10.1.** Для проверки эффективности новой технологии отобраны две группы рабочих: в первой группе численностью $n_1 = 50$ чел., где применялась новая технология, выборочная средняя выработка составила $\bar{x} = 85$ (изделий), во второй группе численностью $n_2 = 70$ чел. выборочная средняя — $\bar{y} = 78$ (изделий). Предварительно установлено, что дисперсии выработки в группах равны соответственно $\sigma_x^2 = 100$ и $\sigma_y^2 = 74$. На уровне значимости $\alpha = 0{,}05$ выяснить влияние новой технологии на среднюю производительность.

Р е ш е н и е. Проверяемая гипотеза H_0: $\bar{x}_0 = \bar{y}_0$, т.е. средние выработки рабочих одинаковы по новой и старой технологиям. В качестве конкурирующей гипотезы можно взять H_1: $\bar{x}_0 > \bar{y}_0$ или H_2: $\bar{x}_0 \ne \bar{y}_0$ (в данной задаче более естественна гипотеза H_1, так как ее справедливость означает эффективность применения новой технологии).

По (**10.5**) фактическое значение статистики критерия

$$t = \frac{85-78}{\sqrt{\frac{100}{50}+\frac{74}{70}}} = 4{,}00.$$

При конкурирующей гипотезе H_1 критическое значение статистики находится из условия (**10.6**), т.е. $\Phi(t_{кр}) = 1 - 2\cdot 0{,}05 = 0{,}9$, откуда по табл. II приложений $t_{кр} = t_{0,9} = 1{,}64$, а при конкурирующей гипотезе H_2 — из условия (**10.7**), т.е. $\Phi(t_{кр}) = 1 - 0{,}05 = 0{,}95$, откуда по таблице $t_{кр} = t_{0,95} = 1{,}96$.

Так как фактически наблюдаемое значение $t = 4{,}00$ больше критического значения $t_{кр}$ (при любой из взятых конкурирующих гипотез), то гипотеза H_0 отвергается, т.е. на 5%-ном уровне значимости можно сделать вывод, что новая технология позволяет повысить среднюю выработку рабочих. ◀

Будем теперь предполагать, что распределение признака (случайной величины) X и Y в каждой совокупности имеет **нормальный закон**. В этом случае, если дисперсии σ_x^2 и σ_y^2 известны, то проверка гипотезы проводится так же, как описано выше, не только для больших, но и для малых по объему выборок.

Если же **дисперсии σ_x^2 и σ_y^2 неизвестны, но равны**, т.е. $\sigma_x^2 = \sigma_y^2 = \sigma^2$, то в качестве неизвестной величины σ^2 можно взять ее оценку — «исправленную» выборочную дисперсию

$$\hat{s}_x^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{или} \quad \hat{s}_y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Однако «лучшей» оценкой для σ^2 будет дисперсия «смешанной» совокупности объема n_1+n_2, т.е.

$$\hat{s}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2}{n_1 + n_2 - 2} = \frac{(n_1-1)\hat{s}_x^2 + (n_2-1)\hat{s}_y^2}{n_1 + n_2 - 2} = \frac{n_1 s_x^2 + n_2 s_y^2}{n_1 + n_2 - 2},$$

а оценкой дисперсии разности независимых выборочных средних $\sigma_{\bar{x}-\bar{y}}^2$ —

$$\hat{s}_{\bar{x}-\bar{y}}^2 = \frac{n_1 s_x^2 + n_2 s_y^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

(обращаем внимание на то, что число степеней свободы $k = n_1 + n_2 - 2$ на 2 меньше общего числа наблюдений n_1+n_2, так как две степени свободы «теряются» при определении по выборочным данным средних $\bar{x}$ и $\bar{y}$).

Доказано, что в случае справедливости гипотезы H_0 статистика

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{n_1 s_x^2 + n_2 s_y^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \textbf{(10.8)}$$

имеет t-распределение Стьюдента с $k = n_1 + n_2 - 2$ степенями свободы. Поэтому критическое значение статистики t находится по тем же формулам (**10.6**) или (**10.7**) в зависимости от типа критической области, в которых вместо функции Лапласа $\Phi(t)$ берется

функция $\theta(t,k)$ для распределения Стьюдента при числе степеней свободы $k=n_1+n_2-2$, т.е. $\theta(t,k)=1-2\alpha$ или $\theta(t,k)=1-\alpha$.

При этом сохраняется то же правило опровержения (принятия) гипотезы: гипотеза H_0 отвергается на уровне значимости α, если $|t|>t_{1-2\alpha;k}$ (в случае односторонней критической области), либо если $|t|>t_{1-\alpha;k}$ (в случае двусторонней критической области); в противном случае гипотеза H_0 не отвергается (принимается).

Замечание. Если дисперсии σ_x^2 и σ_y^2 неизвестны и не предполагается, что они равны, то статистика $t=(\bar{x}-\bar{y})/\hat{s}_{\bar{x}-\bar{y}}$ также имеет t-распределение Стьюдента, однако соответствующее ему число степеней свободы определяется приближенно и более сложным образом.

▷ **Пример 10.2.** Произведены две выборки урожая пшеницы: при своевременной уборке урожая и уборке с некоторым опозданием. В первом случае при наблюдении 8 участков выборочная средняя урожайность составила 16,2 ц/га, а среднее квадратическое отклонение — 3,2 ц/га; во втором случае при наблюдении 9 участков те же характеристики равнялись соответственно 13,9 ц/га и 2,1 ц/га. На уровне значимости $\alpha=0{,}05$ выяснить влияние своевременности уборки урожая на среднее значение урожайности.

Р е ш е н и е. Проверяемая гипотеза H_0: $\bar{x}_0=\bar{y}_0$, т.е. средние значения урожайности при своевременной уборке урожая и с некоторым опозданием равны. В качестве альтернативной гипотезы берем гипотезу H_1: $\bar{x}_0>\bar{y}_0$, принятие которой означает существенное влияние на урожайность сроков уборки.

Фактически наблюдаемое значение статистики критерия по **(10.8)**

$$t=\frac{16{,}2-13{,}9}{\sqrt{\frac{9\cdot 3{,}2^2+8\cdot 2{,}1^2}{8+9-2}\left(\frac{1}{8}+\frac{1}{9}\right)}}=1{,}62.$$

Критическое значение статистики для односторонней области определяется при числе степеней свободы $k=n_1+n_2-2=9+8-2=15$ из условия $\theta(t,k)=1-2\cdot 0{,}05=0{,}9$, откуда по табл. IV приложений $t_{0,9;15}=1{,}75$. Так как $t=1{,}62<t_{0,9;15}=1{,}75$, то гипотеза H_0 принимается. Это означает, что имеющиеся выборочные

данные на 5%-ном уровне значимости не позволяют считать, что некоторое запаздывание в сроках уборки оказывает существенное влияние на величину урожая. Еще раз подчеркнем, что это не означает безоговорочную верность гипотезы H_0. Вполне возможно, что только незначительный объем выборки позволил принять эту гипотезу, а при увеличении объемов выборки (числа отобранных участков) гипотеза H_0 будет отвергнута. ▶

Сравнение средних нескольких совокупностей. Эта задача рассматривается в гл. 11 «Дисперсионный анализ».

Исключение грубых ошибок наблюдений. Рассмотренный критерий можно применять для исключения грубых ошибок наблюдений. Грубые ошибки могут возникнуть из-за ошибок показаний измерительных приборов, ошибок регистрации, случайного сдвига запятой в десятичной записи числа и т.д.

Пусть, например, $x^*, x_1, x_2, \ldots, x_n$ — совокупность имеющихся наблюдений, причем x^* резко выделяется. Необходимо решить вопрос о принадлежности резко выделяющегося значения к остальным наблюдениям.

Для ряда наблюдений $x_1, x_2, \ldots, x_n$ рассчитывают среднюю арифметическую $\bar{x}$ и «исправленное» среднее квадратическое отклонение $\hat{s}$. При справедливости гипотезы H_0: $\bar{x}_0=x^*$ о принадлежности x^* к остальным наблюдениям статистика $t=\dfrac{\bar{x}-x^*}{\hat{s}}$ (получаемая как частный случай из **(10.8)** при $\bar{y}=x^*$, $n_2=1$) имеет t-распределение Стьюдента с $k=n-1$ степенями свободы. Конкурирующая гипотеза H_1 имеет вид: $\bar{x}_0>x^*$ или $\bar{x}_0<x^*$ — в зависимости от того, является ли резко выделяющееся значение больше или меньше остальных наблюдений. Гипотеза H_0 отвергается, если $|t|>t_{\text{кр}}$, и принимается, если $|t|\le t_{\text{кр}}$.

▷ **Пример 10.3.** Имеются следующие данные об урожайности пшеницы на 8 опытных участках одинакового размера (ц/га): 26,5; 26,2; 35,9; 30,1; 32,3; 29,3; 26,1; 25,0. Есть основание предполагать, что значение урожайности третьего участка $x^*=35{,}9$ зарегистрировано неверно. Является ли это значение аномальным (резко выделяющимся) на 5%-ном уровне значимости?

Р е ш е н и е. Исключив значение $x^* = 35{,}9$, найдем для оставшихся наблюдений $\bar{x} = 27{,}93$(ц/га) и $s = 2{,}67$(ц/га). Фактически наблюдаемое значение $t = \dfrac{35{,}9 - 27{,}93}{2{,}67} = 2{,}98$ больше табличного $t_{кр} = t_{1-2\alpha;n-1} = t_{0{,}9;6} = 1{,}94$, следовательно, значение $x^* = 35{,}9$ является аномальным, и его следует отбросить. ▶

10.4. Проверка гипотез о равенстве долей признака в двух и более совокупностях

Сравнение долей признака в двух совокупностях — достаточно часто встречающаяся на практике задача. Например, если выборочная доля признака в одной совокупности отличается от такой же доли в другой совокупности, то указывает ли это на то, что наличие признака в одной совокупности действительно вероятнее, или полученное расхождение долей является случайным?

Сформулируем задачу. Имеются две совокупности, генеральные доли признака в которых равны соответственно p_1 и p_2. Необходимо проверить нулевую гипотезу о равенстве генеральных долей, т.е. H_0: $p_1 = p_2$. Для проверки гипотезы H_0 из этих совокупностей взяты две независимые выборки достаточно большого объема[1] n_1 и n_2. Выборочные доли признака равны соответственно $w_1 = \dfrac{m_1}{n_1}$ и $w_2 = \dfrac{m_2}{n_2}$, где m_1 и m_2 — соответственно число элементов первой и второй выборок, обладающих данным признаком.

При достаточно больших n_1 и n_2, как отмечено в § 9.5, выборочные доли w_1 и w_2 имеют приближенно нормальный закон распределения с математическими ожиданиями p_1 и p_2 и дисперсиями $\dfrac{p_1(1-p_1)}{n_1}$ и $\dfrac{p_2(1-p_2)}{n_2}$, т.е. соответственно $N\left(p_1; \dfrac{p_1(1-p_1)}{n_1}\right)$ и $N\left(p_2; \dfrac{p_2(1-p_2)}{n_2}\right)$. При справедливости гипотезы H_0: $p_1 = p_2 = p$ разность $w_1 - w_2$ имеет нормальный закон распределения с математическим ожиданием $M(w_1 - w_2) = p - p = 0$ и дисперсией $\sigma^2_{w_1 - w_2} = \sigma^2_{w_1} + \sigma^2_{w_2} = p(1-p)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)$. Поэтому статистика

$$t = \frac{w_1 - w_2}{\sigma_{w_1 - w_2}} = \frac{w_1 - w_2}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \qquad \textbf{(10.9)}$$

имеет стандартное нормальное распределение $N(0;1)$.

В качестве неизвестного значения p, входящего в выражение статистики t, берут ее наилучшую оценку $\hat{p}$, равную выборочной доле признака, если две выборки смешать в одну, т.е.

$$\hat{p} = \frac{m_1 + m_2}{n_1 + n_2}. \qquad \textbf{(10.10)}$$

Выбор типа критической области и проверка гипотезы H_0 осуществляются так же, как и выше, при проверке гипотезы о равенстве средних.

▷ **Пример 10.4.** Контрольную работу по высшей математике по индивидуальным вариантам выполняли студенты двух групп первого курса. В первой группе было предложено 105 задач, из которых верно решено 60, во второй группе из 140 предложенных задач верно решено 69. На уровне значимости 0,02 проверить гипотезу об отсутствии существенных различий в усвоении учебного материала студентами обеих групп.

Р е ш е н и е. Имеем гипотезу H_0: $p_1 = p_2 = p$, т.е. доли решенных задач студентами первой и второй групп равны. В качестве альтернативной возьмем гипотезу H_1: $p_1 \neq p_2$.

При справедливости гипотезы H_0 наилучшей оценкой p будет в соответствии с **(10.10)** $\hat{p} = \dfrac{60+69}{105+140} = \dfrac{129}{245} = 0{,}527$. Выборочные доли решенных задач для каждой группы $w_1 = \dfrac{m_1}{n_1} = \dfrac{60}{105} = 0{,}571$ и $w_2 = \dfrac{m_2}{n_2} = \dfrac{69}{140} = 0{,}493$. Статистика критерия по **(10.9)**

1 Здесь ограничиваемся рассмотрением случая больших по объему выборок.

$$t=\frac{0{,}571-0{,}493}{\sqrt{0{,}527(1-0{,}527)\left(\frac{1}{105}+\frac{1}{140}\right)}}=1{,}21.$$

При конкурирующей гипотезе H_1 выбираем критическую двустороннюю область, границы которой определяем из условия (10.7): $\Phi(t_{кр})=1-0{,}02=0{,}98$, откуда по табл. II приложений $t_{кр}=t_{0,98}=2{,}33$. Фактическое значение критерия меньше критического, т.е. $t<t_{0,98}$, следовательно, гипотеза H_0 принимается, т.е. полученные данные не противоречат гипотезе об одинаковом уровне усвоения учебного материала студентами обеих групп. ▶

Сравнение долей признака в нескольких совокупностях. Пусть имеется l совокупностей, генеральные доли которых равны соответственно $p_1,p_2,...,p_l$. Необходимо проверить нулевую гипотезу о равенстве генеральных долей, т.е. H_0: $p_1=p_2=...=p_n=p$ или H_0: $p_i=p$ ($i=1,2,...,l$). Для проверки гипотезы H_0 из этих совокупностей отобраны l независимых выборок достаточно больших объемов $n_1,n_2,...,n_l$. Выборочные доли признака равны соответственно $w_1=m_1/n_1$, $w_2=m_2/n_2$, ..., $w_l=m_l/n_l$, где m_i — число элементов i-й выборки ($i=1,2,...,l$), обладающих данным признаком.

Можно показать, что при справедливости гипотезы H_0 и при $n\to\infty$ статистика

$$\chi^2=\frac{1}{\hat{p}(1-\hat{p})}\sum_{i=1}^{l}n_i(w_i-\hat{p})^2 \qquad (10.11)$$

имеет χ^2-распределение с $l-1$ степенями свободы.

В качестве неизвестного значения $\hat{p}$, входящего в выражение (10.11), берут наилучшую оценку для p, равную выборочной доле признака, если все l выборок смешать в одну, т.е.

$$\hat{p}=\frac{\sum_{i=1}^{l}m_i}{\sum_{i=1}^{l}n_i}. \qquad (10.12)$$

Для проверки гипотезы H_0 обычно берут правостороннюю критическую область. Гипотеза H_0 отвергается, если $\chi^2>\chi^2_{\alpha;l-1}$, где $\chi^2_{\alpha;l-1}$ — критическое значение критерия χ^2, определяемое на уровне значимости α при числе степеней свободы $l-1$.

▷ **Пример 10.5.** По условию примера **10.4** на уровне значимости $\alpha=0{,}05$ выяснить, можно ли считать, что различия в усвоении учебного материала студентами четырех групп первого курса существенны. Дополнительные условия: для третьей группы $m_3=63$, $n_3=125$, для четвертой группы $m_4=105$, $n_4=160$.

Р е ш е н и е. Выдвигаем гипотезу H_0: $p_1=p_2=p_3=p_4=p$ или $p_i=p$ ($i=1,2,3,4$), т.е. доли решенных задач всех групп равны.

Вычислим по формуле (10.12) оценку $\hat{p}$:

$$\hat{p}=\frac{60+65+63+105}{105+140+125+160}=0{,}553.$$

Выборочные доли решенных задач для каждой группы: $w_1=0{,}571$, $w_2=0{,}499$ (см. пример **10.4**), $w_3=63/125=0{,}504$, $w_4=105/160=0{,}656$.

Статистика критерия по (10.11)

$$\chi^2=\frac{1}{0{,}553(1-0{,}553)}\Big[105(0{,}571-0{,}553)^2+140(0{,}499-0{,}553)^2+$$
$$+125(0{,}504-0{,}553)^2+160(0{,}656-0{,}553)^2\Big]=9{,}87.$$

По табл. V приложений $\chi^2_{0,05;3}=7{,}82$. Так как $\chi^2>\chi^2_{0,05;3}$ $(9{,}87>7{,}82)$, то гипотеза H_0 отвергается, т.е. различие в усвоении учебного материала студентами четырех групп значимо или существенно на уровне $\alpha=0{,}05$. ▶

10.5. Проверка гипотез о равенстве дисперсий двух и более совокупностей

Сравнение дисперсий двух совокупностей. Гипотезы о дисперсиях возникают довольно часто, так как дисперсия характеризует такие исключительно важные показатели, как точность машин, приборов, технологических процессов, степень однородности совокупностей, риск, связанный с отклонением доходности активов от ожидаемого уровня, и т.д.

Сформулируем задачу. Пусть имеются две нормально распределенные совокупности, дисперсии которых равны σ_1^2 и σ_2^2. Необходимо проверить нулевую гипотезу о равенстве диспер-

сий, т.е. H_0: $\sigma_1^2 = \sigma_2^2$ относительно конкурирующей H_1: $\sigma_1^2 > \sigma_2^2$ или H_1': $\sigma_1^2 \neq \sigma_2^2$.

Для проверки гипотезы H_0 из этих совокупностей взяты две независимые выборки объемом n_1 и n_2. Для оценки дисперсий σ_1^2 и σ_2^2 используются «исправленные» выборочные дисперсии $\hat{s}_1^2$ и $\hat{s}_2^2$.

Следовательно, задача проверки гипотезы сводится к сравнению дисперсий $\hat{s}_1^2$ и $\hat{s}_2^2$.

При справедливости гипотезы H_0: $\sigma_1^2 = \sigma_2^2 = \sigma^2$ в качестве оценки σ^2 можно взять те же дисперсии $\hat{s}_1^2$ и $\hat{s}_2^2$, рассчитанные по элементам первой и второй выборок.

Напомним (см. § 9.7), что выборочные характеристики $\frac{(n_1-1)\hat{s}_1^2}{\sigma^2}$ и $\frac{(n_2-1)\hat{s}_2^2}{\sigma^2}$ имеют распределение χ^2 соответственно с $k_1=n_1-1$ и $k_2=n_2-1$ степенями свободы, а их отношение $\frac{\frac{1}{k_1}\chi^2(k_1)}{\frac{1}{k_2}\chi^2(k_2)}$ имеет F-распределение Фишера—Снедекора с k_1 и k_2 степенями свободы (см. § 4.9). Следовательно, случайная величина F, определяемая отношением:

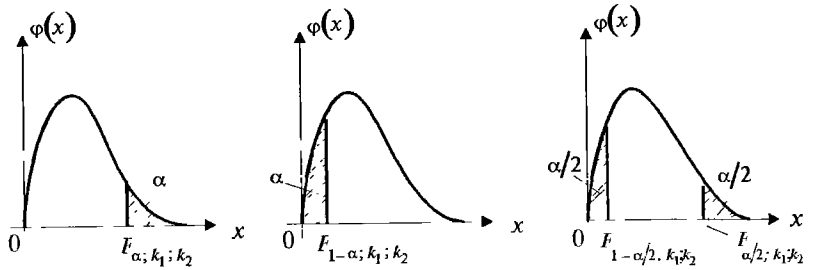
$$F = \frac{\frac{1}{n_1-1}\left[(n_1-1)\hat{s}_1^2/\sigma^2\right]}{\frac{1}{n_2-1}\left[(n_2-1)\hat{s}_2^2/\sigma^2\right]} = \frac{\hat{s}_1^2}{\hat{s}_2^2}, \qquad \textbf{(10.13)}$$

т.е. отношением «исправленных» выборочных дисперсий, имеет *F-распределение Фишера—Снедекора с $k_1=n_1-1$ и $k_2=n_2-1$* степенями свободы. Вид некоторых кривых F-распределения показан на рис. 4.18, а также на рис. 10.5.

При формировании критерия отклонения (принятия) гипотезы H_0 следует учесть, что распределение статистики F (в отличие от нормального или распределения Стьюдента) является несимметричным.

Поэтому гипотеза H_0 отвергается, если $F > F_{\alpha;k_1;k_2}$ (в случае правосторонней критической области — рис. 10.5*а*), либо если $F < F_{1-\alpha;k_1;k_2}$ (в случае левосторонней — рис. 10.5*б*), либо если

$F < F_{1-\alpha/2;k_1;k_2}$ или $F > F_{\alpha/2;k_1;k_2}$ (в случае двусторонней критической области — рис.10.5*в*). В противном случае гипотеза H_0 не отвергается (принимается).



а) б) в)

Рис. 10.5

В приложении VI приведены таблицы значений $F_{\alpha;k_1;k_2}$ для $\alpha = 0{,}05$ и $\alpha = 0{,}01$.

▷ **Пример 10.6.** На двух токарных станках обрабатываются втулки. Отобраны две пробы: из втулок, сделанных на первом станке, n_1=15 шт., на втором станке — n_2=18 шт. По данным этих выборок рассчитаны выборочные дисперсии $s_1^2 = 8{,}5$ (для первого станка) и $s_2^2 = 6{,}3$ (для второго станка). Полагая, что размеры втулок подчиняются нормальному закону распределения, на уровне значимости $\alpha = 0{,}05$ выяснить, можно ли считать, что станки обладают различной точностью.

Р е ш е н и е. Имеем нулевую гипотезу H_0: $\sigma_1^2 = \sigma_2^2$, т.е. дисперсии размера втулок, обрабатываемых на каждом станке, равны. Возьмем в качестве конкурирующей гипотезу H_1: $\sigma_1^2 > \sigma_2^2$ (дисперсия больше для первого станка). Статистика критерия по **(10.13)** (в качестве дисперсии s_1^2, стоящей в числителе, берут большую из двух дисперсий — это дает возможность, учитывая свойства F-распределения, в два раза сократить объем его табличных значений):

$$F = \frac{\hat{s}_1^2}{\hat{s}_2^2} = \frac{\frac{n_1}{n_1-1}s_1^2}{\frac{n_2}{n_2-1}s_2^2} = \frac{(15/14)\cdot 8{,}5}{(18/17)\cdot 6{,}3} = 1{,}37.$$

По табл. VI приложений критическое значение F-критерия на уровне значимости $\alpha = 0{,}05$ при числе степеней свободы $k_1 = n_1 - 1 = 14$ и $k_2 = n_2 - 1 = 17$, т.е. $F_{0,05;14;17} = 2{,}33$. Так как $F < F_{0,05;14;17}$, то гипотеза H_0 не отвергается, т.е. имеющиеся данные не позволяют считать, что станки обладают различной точностью.

Замечание. Если в качестве конкурирующей гипотезы в данной задаче взять гипотезу H_1: $\sigma_1^2 \neq \sigma_2^2$, то, как уже отмечено выше (см. рис. 10.5*в*), следовало взять двустороннюю критическую область и найти $F_{1-\alpha/2;k_1;k_2}$ и $F_{\alpha/2;k_1;k_2}$ соответственно из условий $P\left(F < F_{1-\alpha/2;k_1;k_2}\right) = \dfrac{\alpha}{2}$ и $P\left(F > F_{\alpha/2;k_1;k_2}\right) = \dfrac{\alpha}{2}$. При этом гипотеза H_0 отвергается, если полученное значение $F < F_{1-\alpha/2;k_1;k_2}$ или $F > F_{\alpha/2;k_1;k_2}$

Однако непосредственно по таблицам F-критерия можно найти лишь правую границу $F_{\alpha/2;k_1;k_2}$ (бóльшую единицы), левую же границу $F_{1-\alpha/2;k_1;k_2}$ (меньшую единицы) находят из соотношения, доказанного для F-критерия:

$$F_{1-\alpha/2;k_1;k_2} = \frac{1}{F_{\alpha/2;k_2;k_1}}.$$

В данном случае при $\alpha = 0{,}05$ в задаче следовало найти

$$F_{0,025;14;17} \text{ и } F_{0,975;14;17} = \frac{1}{F_{0,025;17;14}}. \blacktriangleright$$

На практике обычно используется таблица значений F-критерия (см. табл. VI приложений), в которой приведены значения $F_{0.05;k_1;k_2}$ и $F_{0,01;k_1;k_2}$. Это позволяет осуществлять проверку гипотезы H_0 на 5%-ном и 1%-ном уровнях значимости при использовании односторонней критической области, и на 10%-ном и 2%-ном уровнях значимости при двусторонней критической области.

Сравнение дисперсий нескольких совокупностей. Пусть имеется l нормально распределенных совокупностей, дисперсии которых равны соответственно $\sigma_1^2, \sigma_2^2, \ldots, \sigma_l^2$, и l независимых выборок из каждой совокупности объемов $n_1, n_2, \ldots, n_l$. Необходимо проверить нулевую гипотезу о равенстве дисперсий, т.е. H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_l^2 = \sigma^2$ или H_0: $\sigma_i^2 = \sigma^2$ $(i = 1,2,\ldots,l)$.

Для проверки гипотезы H_0 может быть использован *критерий Бартлетта*.

Доказано, что при справедливости гипотезы H_0 и при условии, что $n_i \geq 3 (i = 1,2,\ldots,l)$ статистика

$$\chi^2 = \frac{\sum_{i=1}^{l} (n_i - 1) \ln\left(\overline{s^2} / \hat{s}^2\right)}{1 + \dfrac{1}{3(l-1)}\left(\sum_{i=1}^{l} \dfrac{1}{n_i - 1} - \dfrac{1}{n_1 + \ldots + n_l - l}\right)} \qquad \textbf{(10.13')}$$

(в которой

$$\hat{s}^2 = \frac{n_i s_i^2}{n_i - 1} \qquad \textbf{(10.14)}$$

— исправленная выборочная дисперсия i-й выборки,

$$\overline{s^2} = \frac{1}{n_1 + \ldots + n_l - l} \sum_{i=1}^{l} n_i s_i^2 \qquad \textbf{(10.15)}$$

— оценка средней арифметической дисперсий) имеет χ^2-распределение с $l - 1$ степенями свободы. Поэтому гипотеза H_0 отвергается, если фактически наблюдаемое значение $\chi^2 > \chi^2_{\alpha,l-1}$, где $\chi^2_{\alpha,l-1}$ — критическое значение критерия χ^2, найденное на уровне значимости α при числе степеней свободы $l - 1$.

▷ **Пример 10.7.** По условию примера **10.5** на уровне значимости $\alpha = 0{,}05$ выяснить, можно ли считать, что станки обладают различной точностью, если имеются 4 токарных станка и отобраны соответственно четыре пробы объемов: $n_1 = 15$; $n_2 = 18$; $n_3 = 25$; $n_4 = 32$, а выборочные дисперсии размеров втулок равны соответственно: $s_1^2 = 8{,}5$; $s_2^2 = 6{,}3$; $s_3^2 = 9{,}3$; $s_4^2 = 5{,}8$.

Р е ш е н и е. Имеем нулевую гипотезу H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2$ или $\sigma_i^2 = \sigma^2 (i = 1, 2, 3, 4)$.

По формуле **(10.14)** найдем исправленные выборочные дисперсии размеров втулок:

$$\hat{s}_1^2 = \frac{15}{14} \cdot 8{,}5 = 9{,}11; \qquad \hat{s}_2^2 = \frac{18}{17} \cdot 6{,}3 = 6{,}67;$$

$$\hat{s}_3^2 = \frac{25}{24} \cdot 9{,}3 = 9{,}69; \qquad \hat{s}_4^2 = \frac{32}{31} \cdot 5{,}8 = 5{,}99,$$

а по формуле (**10.15**) — оценку средней арифметической дисперсий

$$\overline{s^2} = \frac{15 \cdot 8{,}5 + 18 \cdot 6{,}3 + 25 \cdot 9{,}3 + 32 \cdot 5{,}8}{15 + 18 + 25 + 32 - 4} = \frac{659}{86} = 7{,}66.$$

Статистика критерия по формуле (**10.13′**) равна:

$$\chi^2 = \frac{14\ln(7{,}66/9{,}11) + 17\ln(7{,}66/6{,}67) + 24\ln(7{,}66/9{,}69) + 31\ln(7{,}66/5{,}99)}{1 + \frac{1}{3 \cdot 3}\left(\frac{1}{14} + \frac{1}{17} + \frac{1}{24} + \frac{1}{31} - \frac{1}{76}\right)} = 1{,}87.$$

По табл. V приложений $\chi^2_{0,05;3} = 7{,}82$.

Так как $\chi^2 < \chi^2_{0,05;3}\,(1{,}87 < 7{,}82)$, то гипотеза H_0 не отвергается, т.е. имеющиеся данные не позволяют считать, что рассматриваемые станки обладают различной точностью. ▶

10.6. Проверка гипотез о числовых значениях параметров

Гипотезы о числовых значениях встречаются в различных задачах. Пусть x_i ($i = 1,2,...,n$) — значения некоторого параметра изделий, производящихся станком автоматической линии, и пусть a — заданное номинальное значение этого параметра. Каждое отдельное значение x_i может, естественно, как-то отклоняться от заданного номинала. Очевидно, для того, чтобы проверить правильность настройки этого станка, надо убедиться в том, что среднее значение параметра у производимых на нем изделий будет соответствовать номиналу, т.е. проверить гипотезу H_0: $\bar{x}_0 = a$ против альтернативной H_1: $\bar{x}_0 \neq a$, или H_2': $\bar{x}_0 < a$, или H_2'': $\bar{x}_0 > a$.

При произвольной настройке станка может возникнуть необходимость проверки гипотезы о том, что точность изготовления изделий по данному параметру, задаваемая дисперсий σ^2, равна заданной величине σ_0^2, т.е. H_0: $\sigma^2 = \sigma_0^2$ или, например, того, что доля бракованных изделий, производимых станком, равна заданной величине p_0, т.е. H_0: $p = p_0$ и т.д.

Аналогичные задачи могут возникнуть, например, в финансовом анализе, когда по данным выборки надо установить, можно ли считать доходность актива определенного вида или

портфеля ценных бумаг, либо ее риск равным заданному числу; или по результатам выборочной аудиторской проверки однотипных документов нужно убедиться, можно ли считать процент допущенных ошибок равным номиналу, и т.п.

В общем случае гипотезы подобного типа имеют вид H_0: $\theta = \Delta_0$, где θ — некоторый параметр исследуемого распределения, а Δ_0 — область его конкретных значений, состоящая в частном случае из одного значения.

При проверке гипотезы указанного типа можно использовать тот же подход, что и в § 10.2 (см., например, проверку гипотезы H_0: $a = a_0$ против альтернативной H_1: $a = a_1 > a_0$ при известной дисперсии σ^2 в примере **10.0**).

Соответствующие критерии проверки гипотез о числовых значениях параметров нормального закона приведены в табл. 10.2.

Таблица 10.2

Нулевая гипотеза	*Предположения*	*Статистика критерия*	*Альтернативная гипотеза*	*Критерий отклонения гипотезы*
$a=a_0$	σ^2 известна	$t=\frac{\bar{x}-a_0}{\sigma/\sqrt{n}}$	$a=a_1>a_0$ $a=a_1<a_0$	$\lvert t\rvert > t_{1-2\alpha}$
			$a=a_1\neq a_0$	$\lvert t\rvert > t_{1-\alpha}$
	σ^2 неизвестна	$t=\frac{\bar{x}-a_0}{s/\sqrt{n-1}}$	$a=a_1>a_0$ $a=a_1<a_0$	$\lvert t\rvert > t_{1-2\alpha,n-1}$
			$a=a_1\neq a_0$	$\lvert t\rvert > t_{1-\alpha,n-1}$
$\sigma^2=\sigma_0^2$	a неизвестно	$\chi^2=\frac{ns^2}{\sigma_0^2}$	$\sigma^2=\sigma_1^2>\sigma_0^2$	$\chi^2>\chi^2_{\alpha;n-1}$
			$\sigma^2=\sigma_1^2<\sigma_0^2$	$\chi^2<\chi^2_{1-\alpha;n-1}$
			$\sigma^2=\sigma_1^2\neq\sigma_0^2$	$\chi^2>\chi^2_{\alpha/2;n-1}$ либо $\chi^2<\chi^2_{1-\alpha/2;n-1}$
$p=p_0$	Достаточно большие n	$t=\frac{w-p_0}{\sqrt{p_0q_0/n}}$	$p=p_1>p_0$ $p=p_1<p_0$	$\lvert t\rvert > t_{1-2\alpha}$
			$p=p_1\neq p_0$	$\lvert t\rvert > t_{1-\alpha}$

Примечание. Критические значения статистик на уровне значимости α определяют по соответствующим таблицам приложений исходя из соотношений:

$$P(|t| < t_{1-\alpha}) = \Phi(t_{1-\alpha}) = 1-\alpha;\ P(|t| < t_{1-\alpha,n-1}) = \theta(t_{1-\alpha,n-1}) = 1-\alpha,$$

$$P(\chi^2 > \chi^2_{\alpha,n-1}) = \alpha.$$

▷ **Пример 10.8.** На основании сделанного прогноза средняя дебиторская задолженность однотипных предприятий региона должна составить a_0=120 ден. ед. Выборочная проверка 10 предприятий дала среднюю задолженность $\bar{x}=135$ ден. ед., а среднее квадратическое отклонение задолженности $s=20$ ден. ед. На уровне значимости 0,05: а) выяснить, можно ли принять данный прогноз; б) найти мощность критерия, использованного в п.а); в) определить минимальное число предприятий, которое следует проверить, чтобы обеспечить мощность критерия 0,975.

Р е ш е н и е. а) Проверяемая гипотеза H_0: $\bar{x}_0=a_0=120$. В качестве альтернативной возьмем гипотезу $H_1:\bar{x}_0=a_1=135$. Так как генеральная дисперсия σ^2 неизвестна, то используем t-критерий Стьюдента. Статистика критерия в соответствии с табл. 10.2 равна $t=\dfrac{\bar{x}-a_0}{s/\sqrt{n-1}}=\dfrac{135-120}{20/\sqrt{10-1}}=2{,}25$. Критическое значение статистики $t_{1-2\cdot 0{,}05;10-1}=t_{0{,}9;9}=1{,}83$.

Так как $|t|>t_{0{,}9;9}\,(2{,}25>1{,}83)$, то гипотеза H_0 отвергается, т.е. на 5%-ном уровне значимости сделанный прогноз должен быть отвергнут.

б) Так как a_1=135>a_0=120, то критическая область правосторонняя и критическое значение выборочной средней

$$\bar{x}_{\text{кр}}=\bar{x}_0+t_{1-2\alpha,n-1}\frac{s}{\sqrt{n-1}}=a+t_{0,9;9}\frac{s}{\sqrt{n-1}}=$$

$$=120+1{,}83\frac{20}{\sqrt{10-1}}=132{,}2\ \text{(ден. ед.)},$$

т.е. критическая область значений для $\bar{x}$ есть интервал (132,2;+∞). Мощность критерия (см. § 10.2) равна вероятности P отвергнуть гипотезу H_0, когда верна гипотеза H_1, т.е.

$$P=P(132{,}2<\bar{x}<+\infty)=\frac{1}{2}-\frac{1}{2}\theta(t,n-1),$$

где $\theta\,(t,\ n-1)$ — функция, выражающая вероятность попадания случайной величины, имеющей t-распределение Стьюдента, на отрезок $(-t,\ t)$ (аналогична функции Лапласа для нормального распределения (см. § 9.7);

$$t=\frac{\bar{x}-a_1}{s/\sqrt{n-1}}=\frac{132{,}2-135}{20/\sqrt{10-1}}=-0{,}42.$$

По табл. IV приложений[1] $\theta(-0{,}42;9)=-\theta(0{,}42;\ 9)\approx -0{,}31$.

Итак, $P=\dfrac{1}{2}-\dfrac{1}{2}\theta(-0{,}42;9)\approx\dfrac{1}{2}(1+0{,}31)\approx 0{,}66$. ▶

в) Воспользуемся решением примера **10.0** *б*), в котором формула (**10.1′**) объема выборки была получена для случая *нормального* закона распределения $\bar{x}$, когда *известна* генеральная дисперсия σ^2. Так как у нас σ^2 не известна, а известна лишь ее выборочная оценка s^2, то статистика критерия $t=\dfrac{\bar{x}-a_0}{s/\sqrt{n-1}}$ имеет t-распределение Стьюдента (см. табл. 10.2), и соответствующая скорректированная формула для n примет вид:

$$n=\frac{\left(t_{1-2\alpha;\,n-1}+t_{1-2\beta;\,n-1}\right)^2 s^2}{(a_1-a_0)^2}$$

При α=0,05, β = 0,025 (ибо по условию мощность критерия 1−β = 0,975), a_0 = 120, a_1 =135, s = 20 получим:

$$n=\frac{16}{9}(t_{0,9;\,n-1}+t_{0,95;\,n-1})^2 \qquad (*)$$

Так как правая часть равенства сама зависит от неизвестного значения n, то n находится приближенно подбором. Так, при n = 20, n = 30, равенство (*) не выполняется (например, при $n=$ 20 $\quad 20\neq\dfrac{16}{9}(t_{0,9;19}+t_{0,95;19})^2=\dfrac{16}{9}(1{,}73+2{,}09)^2=24{,}7)$, а при $n=25\quad 25\approx\dfrac{16}{9}(t_{0,9;24}+t_{0,95;24})^2=\dfrac{16}{9}(1{,}71+2{,}06)^2=25{,}3$.

Следовательно, необходимо проверить 25 предприятий. ▶

Аналогично проверяются и другие гипотезы о числовых значениях параметров в соответствии с критериями проверки, приведенными в табл. 10.2.

[1] Так как непосредственно значений $\theta(t,\ n)$ в данной таблице нет, «внутри» ее в строке k = 9 находим близкие к 0,42 значения 0,40 и 0,54, соответствующие вероятностям γ = 0,3 и γ = 0,4, т.е. θ(0,40; 9) = 0,3, и θ(0,54; 9) = 0,4, а искомое значение θ(0,42; 9) ≈ 0,31 находим интерполированием.

При проверке статистических гипотез есть и другой подход, основанный на том, что выше (в § 9.3) для параметров $\bar{x}_0, p, \sigma^2$ были построены доверительные интервалы. И если параметр $\bar{x}_0$ (или p, или σ^2) не попадает в доверительный интервал с надежностью $\gamma = 1-\alpha$, т.е. попадает в критическую область, то гипотеза H_0 отвергается; в противном случае полагают, что имеющиеся данные не противоречат гипотезе H_0.

Достоинством такого подхода, основанного на построении доверительного интервала для параметра, является то, что кроме проверки гипотезы H_0 получается дополнительная информация о возможных истинных значениях параметра. Однако этот подход применим, если в качестве конкурирующих выступают гипотезы типа $\bar{x}_0 \neq a$, $p \neq p_0$, $\sigma^2 \neq \sigma_0^2$, предполагающие выбор двусторонней критической области.

▷ **Пример 10.9.** По данным примера **9.10** на уровне значимости $\alpha \approx 0{,}05$ проверить гипотезу о том, что средняя выработка рабочих всего цеха равна 121%.

Решение. Проверяемая гипотеза H_0: $\bar{x}_0 = 121(\%)$. Конкурирующая гипотеза H_1: $\bar{x}_0 \neq 121$. В примере **9.10** с надежностью $\gamma \approx 1-0{,}05 = 0{,}95$ построен доверительный интервал для $\bar{x}_0$: $117{,}33 \leq \bar{x}_0 \leq 121{,}07$. Так как значение a=121 принадлежит этому интервалу, то гипотеза H_0 не отвергается, т.е. имеющиеся данные не противоречат предположению о том, что средняя выработка рабочих равна 121%. ▶

▷ **Пример 10.10.** По данным примера **9.11** на уровне значимости α=0,05 проверить гипотезу о том, что доля нестандартных деталей во всей партии равна 12%.

Решение. Проверяемая гипотеза H_0: p=0,12 (или 12%). Конкурирующая гипотеза H_1: $p \neq 0{,}12$. В примере **9.11** с надежностью $\gamma \approx 1-0{,}05 = 0{,}95$ построен доверительный интервал для p: $0{,}044 \leq p \leq 0{,}116$. Так как значение p_0=0,12 не принадлежит этому интервалу, то на уровне значимости α=0,05 гипотеза H_0 отвергается, т.е. имеющиеся данные не позволяют считать, что в партии находится 12% нестандартных деталей. ▶

▷ **Пример 10.11.** По данным примера **9.17** на уровне значимости α=0,1 проверить гипотезу о том, что среднее квадратическое отклонение суточной выработки работниц равно 20 м/ч.

Решение. Проверяемая гипотеза H_0: $\sigma^2 = 20^2 = 400$. Конкурирующая гипотеза H_1: $\sigma^2 \neq 400$. В примере **9.17** с надежностью $\gamma = 1-0{,}1 = 0{,}9$ получен доверительный интервал для σ^2: $157{,}3 \leq \sigma^2 \leq 468{,}9$. Так как значение $\sigma_0^2 = 400$ принадлежит этому интервалу, то на уровне значимости α=0,1 гипотеза H_0 не отвергается, т.е. имеющиеся данные не противоречат предположению о том, что среднее квадратическое отклонение суточной выработки работниц равно 20 м/ч. ▶

10.7. Построение теоретического закона распределения по опытным данным. Проверка гипотез о законе распределения

Одной из важнейших задач математической статистики является ***установление теоретического закона распределения случайной величины***, характеризующей изучаемый признак по опытному (эмпирическому) распределению, представляющему вариационный ряд.

Для решения этой задачи необходимо определить вид и параметры закона распределения.

Предположение о **виде закона распределения** может быть выдвинуто исходя из теоретических предпосылок (например, выполнение условий центральной предельной теоремы может свидетельствовать о нормальном законе распределения случайной величины), опыта аналогичных предшествующих исследований и, наконец, на основании графического изображения эмпирического распределения.

Параметры распределения, как правило, неизвестны, поэтому их заменяют наилучшими оценками по выборке, как это сделано в гл. 9.

Как бы хорошо ни был подобран теоретический закон распределения, между эмпирическим и теоретическим распределениями неизбежны расхождения. Естественно возникает вопрос: объясняются ли эти расхождения только случайными обстоятельствами, связанными с ограниченным числом наблюдений,

или они являются существенными и связаны с тем, что теоретический закон распределения подобран неудачно. Для ответа на этот вопрос и служат **критерии согласия**.

Пусть необходимо проверить нулевую гипотезу H_0 о том, что исследуемая случайная величина X подчиняется определенному закону распределения. Для проверки гипотезы H_0 выбирают некоторую случайную величину U, характеризующую степень расхождения теоретического и эмпирического распределений, закон распределения которой при достаточно больших n известен и практически не зависит от закона распределения случайной величины X.

Зная закон распределения U, можно найти вероятность того, что U приняла значение не меньше, чем фактически наблюдаемое в опыте u, т.е. $U \geq u$. Если $P(U \geq u) = \alpha$ мала, то это означает в соответствии с принципом практической уверенности, что такие, как в опыте, и бóльшие отклонения практически невозможны. В этом случае гипотезу H_0 отвергают. Если же вероятность $P(U \geq u) = \alpha$ не мала, расхождение между эмпирическим и теоретическим распределениями несущественно и гипотезу H_0 можно считать правдоподобной или по крайней мере не противоречащей опытным данным.

χ^2-критерий Пирсона. В наиболее часто используемом на практике *критерии χ^2-Пирсона* в качестве меры расхождения U берется величина χ^2, равная сумме квадратов отклонений частостей (статистических вероятностей) w_i от гипотетических p_i, рассчитанных по предполагаемому распределению, взятых с некоторыми весами c_i:

$$U = \chi^2 = \sum_{i=1}^{m} c_i (w_i - p_i)^2.$$

Веса c_i вводятся таким образом, чтобы при одних и тех же отклонениях $(w_i - p_i)^2$ больший вес имели отклонения, при которых p_i мала, и меньший вес — при которых p_i велика. Очевидно, этого удается достичь, если взять c_i обратно пропорциональными вероятностям p_i. Взяв в качестве весов $c_i = \dfrac{n}{p_i}$, можно доказать, что при $n \to \infty$ статистика

$$U = \chi^2 = \sum_{i=1}^{m} \frac{n}{p_i} (w_i - p_i)^2,$$

или

$$U = \chi^2 = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i} \qquad \textbf{(10.16)}$$

имеет χ^2-распределение с $k = m - r - 1$ степенями свободы, где m — число интервалов эмпирического распределения (вариационного ряда); r — число параметров теоретического распределения, вычисленных по экспериментальным данным.

Числа $n_i = nw_i$ и np_i называются соответственно *эмпирическими* и *теоретическими частотами*.

Схема применения критерия χ^2 для проверки гипотезы H_0 сводится к следующему:

1. Определяется мера расхождения эмпирических и теоретических частот χ^2 по **(10.16)**.

2. Для выбранного уровня значимости α по таблице χ^2-распределения находят критическое значение $\chi^2_{\alpha;k}$ при числе степеней свободы $k = m - r - 1$.

3. Если фактически наблюдаемое значение χ^2 больше критического, т.е. $\chi^2 > \chi^2_{\alpha;k}$, то гипотеза[1] H_0 отвергается, если $\chi^2 \leq \chi^2_{\alpha;k}$, гипотеза H_0 не противоречит опытным данным.

Замечание. Как уже отмечено, статистика

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}$$

имеет χ^2-распределение лишь при $n \to \infty$, поэтому необходимо, чтобы в каждом интервале было достаточное количество наблюдений, по крайней мере 5 наблюдений. Если в каком-нибудь интервале число наблюдений $n_i < 5$, имеет смысл объединить соседние интервалы[2], чтобы в объединенных интервалах n_i было не меньше 5.

▷ **Пример 10.12.** Для эмпирического распределения рабочих цеха по выработке по данным первых двух граф табл. 8.1 подоб-

[1] Так как вероятность $P(\chi^2 > \chi^2_{\alpha;k}) = \alpha$ мала, то выполнение неравенства $\chi^2 > \chi^2_{\alpha;k}$ практически невозможно, если гипотеза H_0 верна.

[2] Поэтому при вычислении числа степеней свободы в качестве величины m берется соответственно уменьшенное число интервалов.

рать соответствующее теоретическое распределение и на уровне значимости $\alpha = 0{,}05$ проверить гипотезу о согласованности двух распределений с помощью критерия χ^2.

Р е ш е н и е. По виду гистограммы распределения рабочих по выработке (рис. 10.6) можно предположить нормальный закон распределения признака. Параметры нормального закона a и σ^2, являющиеся соответственно математическим ожиданием и дисперсией случайной величины X, неизвестны, поэтому заменяем их «наилучшими» оценками по выборке — несмещенными и состоятельными оценками соответственно выборочной средней $\bar{x}$ и «исправленной» выборочной дисперсией $\hat{s}^2$. Так как число наблюдений n=100 достаточно велико, то вместо «исправленной» $\hat{s}^2$ можно взять «обычную» выборочную дисперсию s^2. В примере **8.8** вычислены $\bar{x} = 119{,}2(\%)$, $s^2 = 87{,}48$, $s = 9{,}35(\%)$.

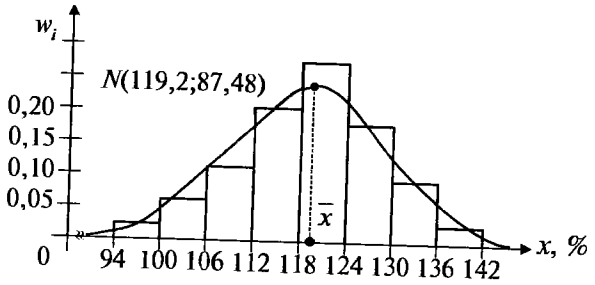


Рис. 10.6

Итак, выдвигаемая гипотеза H_0: случайная величина X — выработка рабочих цеха — распределена нормально с параметрами $a = 119{,}2$; $\sigma^2 = 87{,}48$, т.е. $X \sim N(119{,}2;\ 87{,}48)$.

Для расчета вероятностей p_i попадания случайной величины X в интервал $[x_i, x_{i+1}]$ используем функцию Лапласа в соответствии со свойством нормального распределения:

$$p_i\left(x_i \le X \le x_{i+1}\right) = \frac{1}{2}\left[\Phi\left(\frac{x_{i+1}-a}{\sigma}\right) - \Phi\left(\frac{x_i - a}{\sigma}\right)\right] \approx$$

$$\approx \frac{1}{2}\left[\Phi\left(\frac{x_{i+1}-119{,}2}{9{,}35}\right) - \Phi\left(\frac{x_i - 119{,}2}{9{,}35}\right)\right].$$

Например, $p_1(94 \le X \le 100) = \frac{1}{2}\left[\Phi\left(\frac{100-119{,}2}{9{,}35}\right) - \Phi\left(\frac{94-119{,}2}{9{,}35}\right)\right] =$

$= \frac{1}{2}\left[\Phi(-2{,}05) - \Phi(-2{,}69)\right] = \frac{1}{2}(-0{,}9596 + 0{,}9928) = 0{,}0166$ и соответствующая первому интервалу теоретическая частота $np_1 = 100 \cdot 0{,}0166 \approx 1{,}7$ и т.д.

Для определения статистики χ^2 удобно составить таблицу:

Таблица 10.3

i	Интервал $[x_i, x_{i+1}]$	Эмпирические частоты n_i	Вероятности p_i	Теоретические частоты np_i	$(n_i - np_i)^2$	$\frac{(n_i - np_i)^2}{np_i}$
1	94—100	3 } 10	0,017	1,7 } 7,6	5,76	0,758
2	100—106	7	0,059	5,9		
3	106—112	11	0,141	14,1	9,61	0,682
4	112—118	20	0,228	22,8	7,84	0,344
5	118—124	28	0,247	24,7	10,89	0,441
6	124—130	19	0,182	18,2	0,64	0,035
7	130—136	10 } 12	0,087	8,7 } 11,6	0,16	0,014
8	136—142	2	0,029	2,9		
	$\sum$	100	0,990	99,0	–	$\chi^2 = 2{,}27$

Учитывая, что в рассматриваемом эмпирическом распределении частоты первого и последнего интервалов (n_1=3, n_8=2) меньше 5, при использовании критерия χ^2-Пирсона в соответствии с замечанием на с. 375 целесообразно объединить указанные интервалы с соседними (см. табл. 10.3).

Итак, фактически наблюдаемое значение статистики χ^2=2,27.

Так как новое число интервалов (с учетом объединения крайних) m=6, а нормальный закон распределения определяется $r = 2$ параметрами, то число степеней свободы $k = m - r - 1 = 6 - 2 - 1 = 3$. Соответствующее критическое значение статистики χ^2 по табл. V приложений $\chi^2_{0,05;3.} = 7{,}82$. Так как $\chi^2 < \chi^2_{0,05;3}$, то гипотеза о выбранном теоретическом нормальном законе $N(119{,}2;\ 87{,}48)$ согласуется с опытными данными. ▶

Замечание. Для графического изображения эмпирического и выравнивающего его теоретического нормального распределений необходимо использовать одинаковый для двух распределений масштаб по оси ординат.

Так, если при построении гистограммы эмпирического распределения по оси ординат откладывать *плотность частости* $\frac{n_i}{n\Delta x}$ (где n_i — частота i-го интервала (i=1,2,...,m), Δx — величина интервала, m — число интервалов, n — число наблюдений, объем выборки), то выравнивать такую гистограмму будет теоретическая нормальная кривая с плотностью $\varphi_N(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-a)^2/2\sigma^2}$, где в качестве параметров a и σ^2 используются их состоятельные и несмещенные выборочные оценки: соответственно средняя $\bar{x}$ и дисперсия $\hat{s}^2$ (либо $s^2 \approx \hat{s}^2$ при больших n).

Для построения кривой $\varphi_N(x)$ можно использовать таблицу плотности вероятности стандартного нормального распределения (табл. I приложений) в соответствии с формулой

$$\varphi_N(x)=\frac{1}{\sigma}f(t), \quad \text{где } f(t)=\frac{1}{\sqrt{2\pi}}e^{-t^2/2} \text{ и } t=\frac{x-a}{\sigma}\approx\frac{x-\bar{x}}{s}.$$

При равенстве величин всех интервалов (как в примере **10.12**) часто бывает удобнее при построении гистограммы эмпирического распределения по оси ординат откладывать частости $w_i=\frac{n_i}{n}$ (см. рис. 10.6) или частоты n_i. В этом случае выравнивающей гистограмму кривой будет растянутая (сжатая) вдоль оси ординат в Δx (или $n\Delta x$) раз нормальная кривая, т.е. кривая $\varphi_1(x)=\varphi_N(x)\Delta x$ (или кривая $\varphi_2(x)=\varphi_N(x)n\Delta x$).

Точное построение выравнивающей кривой $\varphi_1(x)$ (или $\varphi_2(x)$) связано с проведением дополнительных расчетов. Их можно избежать, используя приближенный способ построения (рис. 10.6). В процессе применения χ^2-критерия Пирсона были вычислены вероятности p_i и теоретические частоты np_i интервалов распределения. Учитывая, что в соответствии со свойствами плотности распределения $\varphi_N(x_i)\Delta x_i \approx p_i$ (или $n\varphi_N(x_i)\Delta x_i \approx np_i$), выравнивающую теоретическую кривую $\varphi_1(x)$ (или $\varphi_2(x)$) можно построить приближенно по точкам (x_i, p_i) (или (x_i, np_i)), где в качестве значений x_i (i=1,2,...,m) целесообразно взять середины интервалов (рис. 10.6). При этом следует иметь в виду, что максимум выравнивающей кривой $\varphi_1(x)$ (или $\varphi_2(x)$) будет в точке $x = a \approx \bar{x}$ и равен

$$\frac{\Delta x}{\sigma}f(0)\approx 0{,}3989\frac{\Delta x}{s} \quad (\text{или } \frac{n\Delta x}{\sigma}f(0)\approx 0{,}3989\frac{n\Delta x}{s}).$$

▷ **Пример 10.12а.** Имеются следующие статистические данные о числе вызовов специализированных бригад скорой помощи в час в некотором населенном пункте в течение 300 ч:

Число вызовов в час x_i	0	1	2	3	4	5	6	7	8	Σ
Частота n_i	15	71	75	68	39	17	10	4	1	300

Подобрать соответствующее теоретические распределение и на уровне значимости $\alpha = 0{,}05$ поверить гипотезу о согласованности двух распределений с помощью критерия χ^2.

Р е ш е н и е. Вычислим выборочные среднюю и дисперсию:

$$\bar{x}=\frac{\sum_{i=1}^{m}x_i n_i}{n}=\frac{0\cdot 15+\ldots+8\cdot 1}{300}=2{,}54;$$

$$s^2=\overline{x^2}-\bar{x}^2=\frac{\sum_{i=1}^{m}x_i^2 n_i}{n}-x^2=\frac{0^2\cdot 15+\ldots+8^2\cdot 1}{300}-2{,}54^2\approx 2{,}39.$$

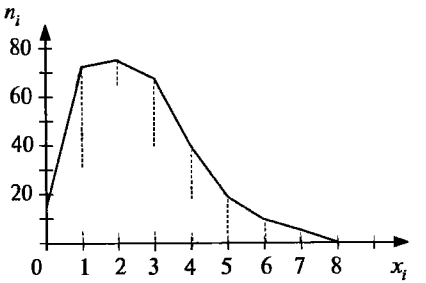


Рис. 10.6а

Выдвигаем гипотезу H_0: случайная величина X — число вызовов скорой помощи в час — распределена по закону Пуассона с параметром $\lambda = 2{,}54$.

В пользу этой гипотезы свидетельствует следующее:

— вызов скорой помощи для каждого жителя — событие в целом достаточно редкое;

— полигон частостей (частот) дискретной случайной величины X (рис. 10.6*а*) по своему виду напоминает полигон пуассоновского распределения вероятностей при небольших значениях λ (см. передний форзац учебника):

— оценки математического ожидания $M(X)$ и дисперсии $D(X)$ — выборочная средняя и выборочная дисперсия приближенно равны, т.е. $\bar{x} \approx s^2$ (а равенство $M(X) = D(X)$, или $a = \sigma^2$, характерно именно для распределения Пуассона — см. § 4.2).

В качестве неизвестного параметра λ, являющегося математическим ожиданием случайной величины, распределенной по закону Пуассона (см. § 4.2), берем его несмещенную и состоятельную оценку по выборке — выборочную среднюю, т.е. $\lambda \approx \bar{x} = 2{,}54$.

Вероятности значений случайной величины X найдем по формуле **(4.8)**:

$$p_i = P(X = x_i = m) = \frac{2{,}54^m e^{-2{,}54}}{m!}.$$

Для определения статистики χ^2 составим таблицу:

Таблица 10.3а

i	$x_i=m$	n_i	p_i	np_i	$(n_i-np_i)^2$	$\frac{(n_i-np_i)^2}{np_i}$
1	0	15	0,0789	23,7	75,69	3,194
2	1	71	0,2003	60,1	98,01	1,631
3	2	75	0,2544	76,3	1,69	0,022
4	3	68	0,2154	64,6	11,56	0,179
5	4	39	0,1368	41,0	3,61	0,088
6	5	17	0,0695	20,9	14,44	0,694
7	6	10	0,0294	8,8	1,44	0,164
8	7	4 } 5	0,0107	3,2 } 4,2	0,64	0,152
9	8	1	0,0034	1,0		
	Σ	300	0,9988	299,6	—	$\chi^2=6{,}12$

При расчете χ^2 объединяем последние два интервала, так как их частоты ($n_8 = 4$, $n_9 = 1$) меньше 5.

Так как новое число интервалов (с учетом объединения двух последних) $m = 8$, а закон Пуассона определяется $r = 1$ параметром, то число степеней свободы $k = m - r - 1 = 8 - 1 - -1 = 6$. По табл. V приложений $\chi^2_{0,05;6} = 12{,}59$. Так как $\chi^2 < \chi^2_{0.05:6}$ $(6{,}12 < 12{,}59)$, то гипотеза H_0 согласуется с опытными данными. ▶

Критерий Колмогорова. На практике кроме критерия χ^2 часто используется критерий Колмогорова, в котором в качестве меры расхождения между теоретическим и эмпирическим распределениями рассматривают максимальное значение абсолютной величины разности между эмпирической функцией распределения $F_n(x)$ и соответствующей теоретической функцией распределения

$$D = \max\left|F_n(x) - F(x)\right|, \qquad \textbf{(10.17)}$$

называемое *статистикой критерия Колмогорова.*

Доказано, что какова бы ни была функция распределения $F(x)$ непрерывной случайной величины X, при неограниченном увеличении числа наблюдений ($n \to \infty$) вероятность неравенства $P\left(D\sqrt{n} \geq \lambda\right)$ стремится к пределу

$$P(\lambda) = 1 - \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2\lambda^2}. \qquad \textbf{(10.18)}$$

Задавая уровень значимости α, из соотношения

$$P(\lambda_\alpha) = \alpha \qquad \textbf{(10.19)}$$

можно найти соответствующее критическое значение λ_α. В табл. 10.4 приводятся критические значения λ_α критерия Колмогорова для некоторых α.

Таблица 10.4

Уровень значимости α	0,40	0,30	0,20	0,10	0,05	0,025	0,01	0,005	0,001	0,0005
Критическое значение λ_α	0,89	0,97	1,07	1,22	1,36	1,48	1,63	1,73	1,95	2,03

Схема применения критерия Колмогорова следующая:

1. Строятся эмпирическая функция распределения $F_n(x)$ и предполагаемая теоретическая функция распределения $F(x)$.

2. Определяется мера расхождения между теоретическим и эмпирическим распределением D по формуле **(10.17)** и вычисляется величина

$$\lambda = D\sqrt{n}. \qquad \textbf{(10.20)}$$

3. Если вычисленное значение λ окажется больше критического λ_α, определенного на уровне значимости α, то нулевая гипотеза H_0 о том, что случайная величина X имеет заданный закон распределения, отвергается. Если $\lambda \leq \lambda_\alpha$, то считают, что гипотеза H_0 не противоречит опытным данным

▷ **Пример 10.13.** По данным примера **10.12** и табл. 8.1 с помощью критерия Колмогорова на уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0 о том, что случайная величина X — выработка рабочих предприятия — имеет нормальный закон распределения с параметрами $a = 119{,}2$; $\sigma^2 = 87{,}48$, т.е. $N(119{,}2;\ 87{,}48)$.

Значения эмпирической функции распределения $F_n(x)$, или накопленной частости, вычислены выше в табл. 8.1, а ее график приведен на рис. 8.2*б* — эти значения и график воспроизводятся соответственно в табл. 10.5 и на рис. 10.7. Для построения теоретической функции распределения для нормального закона воспользуемся ее выражением **(4.30)** через функцию Лапласа:

$$F(x) = \frac{1}{2} + \frac{1}{2}\Phi\left(\frac{x - 119{,}2}{9{,}35}\right).$$

Например, $F(94) = \frac{1}{2} + \frac{1}{2}\Phi\left(\frac{94 - 119{,}2}{9{,}35}\right) = \frac{1}{2} + \frac{1}{2}\Phi(-2{,}69) =$

$= 0{,}5 - 0{,}5 \cdot 0{,}9928 = 0{,}0036 \approx 0{,}004$ и т.д. Результаты вычислений сведем в табл. 10.5, а график $F(x)$ представим на рис. 10.7.

Таблица 10.5

x	94	100	106	112	118	124	130	136	142
$F_n(x)$	0,010	0,030	0,100	0,210	0,410	0,690	0,880	0,980	1,000
$F(x)$	0,004	0,021	0,080	0,221	0,449	0,695	0,878	0,964	0,993

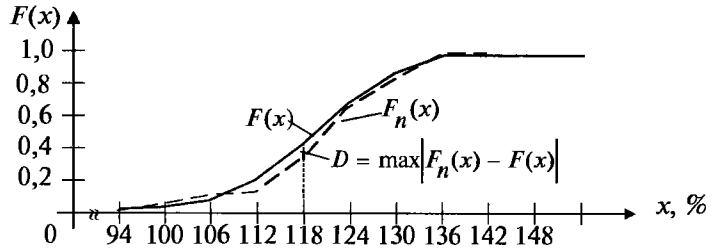


Рис. 10.7

Из рис. 10.7 следует, что

$$D = |F_n(118) - F(118)| = |0{,}410 - 0{,}449| = 0{,}039.$$

По формуле **(10.20)** величина $\lambda = D\sqrt{n} = 0{,}039\sqrt{100} = 0{,}39$.

Критическое значение критерия Колмогорова по табл. 10.4 равно $\lambda_{0{,}05} = 1{,}36$. Так как $\lambda < \lambda_{0{,}05}\ (0{,}39 < 1{,}36)$, то гипотеза H_0 согласуется с опытными данными. ▶

Критерий Колмогорова достаточно часто применяется на практике благодаря своей простоте. Однако в принципе его применение возможно лишь тогда, когда теоретическая функция распределения $F(x)$ задана полностью. Но такой случай на практике встречается весьма редко. Обычно из теоретических соображений известен лишь вид функции распределения, а ее параметры определяются по эмпирическим данным. При применении критерия χ^2 это обстоятельство учитывается соответствующим уменьшением числа степеней свободы. Такого рода поправок в критерии Колмогорова не предусмотрено. Поэтому, если при неизвестных значениях параметров применить критерий Колмогорова, взяв за значения параметров их оценки, то получим завышенное значение вероятности $P(\lambda)$, а значит, бóльшее критическое значение λ_α. В результате есть риск в ряде случаев принять нулевую гипотезу H_0 о законе распределения случайной величины как правдоподобную, в то время как на самом деле она противоречит опытным данным.

10.8. Проверка гипотез об однородности выборок

Гипотезы об *однородности выборок* — это гипотезы о том, что рассматриваемые выборки извлечены из одной и той же генеральной совокупности.

Пусть имеются две независимые выборки, произведенные из генеральных совокупностей с неизвестными теоретическими функциями распределения $F_1(x)$ и $F_2(x)$. Проверяемая нулевая гипотеза имеет вид H_0: $F_1(x) = F_2(x)$ против конкурирующей H_1: $F_1(x) \neq F_2(x)$. Будем предполагать, что функции $F_1(x)$ и $F_2(x)$ непрерывны.

Критерий Колмогорова—Смирнова использует ту же самую идею, что и критерий Колмогорова, но только в критерии Колмогорова сравнивается эмпирическая функция распределения с теоретической, а в критерии Колмогорова—Смирнова сравниваются две эмпирические функции распределения.

Статистика критерия Колмогорова—Смирнова имеет вид:

$$\lambda' = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max\left| F_{n_1}(x) - F_{n_2}(x) \right|, \qquad \textbf{(10.21)}$$

где $F_{n_1}(x)$ и $F_{n_2}(x)$ — эмпирические функции распределения, построенные по двум выборкам объемов n_1 и n_2.

Гипотеза H_0 отвергается, если фактически наблюдаемое значение статистики λ' больше критического $\lambda'_{кр}$, т.е. $\lambda' > \lambda'_{кр}$, и принимается в противном случае.

При малых объемах выборок $(n_1, n_2 \le 20)$ критические значения $\lambda'_{кр}$ для заданных уровней значимости критерия можно найти в специальных таблицах. При $n_1, n_2 \to \infty$ (а практически при $n_1 \ge 50$, $n_2 \ge 50$) распределение статистики λ' сходится к распределению Колмогорова для статистики λ. Поэтому гипотеза H_0 отвергается на уровне значимости α, если фактически наблюдаемое значение λ' больше критического λ_α, т.е. $\lambda' > \lambda_\alpha$, и принимается в противном случае.

▷ **Пример 10.14.** В течение месяца выборочно осуществлялась проверка торговых точек города по продаже овощей. Результаты двух проверок по недовесам покупателям одного вида овощей приведены в табл. 10.6.

Таблица 10.6

Номер интервала	*Интервалы недовесов, г*	*Частоты*	
		n_{i_1} *для выборки* 1	n_{i_2} *для выборки* 2
1	0—10	3	5
2	10—20	10	12
3	20—30	15	8
4	30—40	20	25
5	40—50	12	10
6	50—60	5	8
7	60—70	25	20
8	70—80	15	7
9	80—90	5	5
	Σ	n_1=110	n_2=100

Можно ли считать, что на уровне значимости $\alpha = 0{,}05$ по результатам двух проверок (случайных выборок) недовесы овощей описываются одной и той же функцией распределения?

Р е ш е н и е. Обозначим: $n_{i_1}^{нак}$ и $n_{i_2}^{нак}$ — накопленные частоты соответственно выборок 1 и 2; $F_{n_1}(x_i) = n_{i_1}^{нак}/n_1$, $F_{n_2}(x_i) = n_{i_2}^{нак}/n_2$ — значения их эмпирических функций распределения. Результаты вычислений сведем в табл. 10.7.

Таблица 10.7

x_i	$n_{i_1}^{нак}$	$n_{i_2}^{нак}$	$F_{n_1}(x_i)$	$F_{n_2}(x_i)$	$\left\| F_{n_1}(x_i) - F_{n_2}(x_i) \right\|$
10	3	5	0,027	0,050	0,023
20	13	17	0,118	0,170	0,052
30	28	25	0,254	0,250	0,004
40	48	50	0,436	0,500	0,064
50	60	60	0,545	0,600	0,055
60	65	68	0,591	0,680	**0,089**
70	90	88	0,818	0,880	0,072
80	105	95	0,955	0,950	0,005
90	110	100	1,000	1,000	0,000

Из последнего столбца видно, что $\max\left| F_{n_1}(x_i) - F_{n_2}(x_i) \right| = 0{,}089$.

По формуле **(10.21)** наблюдаемое значение статистики при n_1=110, n_2=100 $\lambda' = \sqrt{\frac{110 \cdot 100}{110 + 100}} \cdot 0{,}089 = 0{,}644$. По табл. 10.4 при $\alpha = 0{,}05 \lambda_{0{,}05} = 1{,}36$.

Так как $\lambda' < \lambda_{0{,}05} (0{,}644 < 1{,}36)$, то нулевая гипотеза H_0 не отвергается, следовательно, недовесы покупателям описываются одной и той же функцией распределения, т.е. они являются устойчивым и закономерным процессом при продаже овощей в данном городе. ▶

Если данные сгруппированы, то для проверки однородности двух или нескольких выборок можно использовать критерий χ^2.

Пусть имеется l независимых выборок объемом n_i (i=1,2,..., l) и данные выборки сгруппированы в m интервалов (групп), а n_{ij} — число элементов j-й выборки, попавшей в i-й интервал.

Проверяется гипотеза H_0 о том, что все l выборок извлечены из одной и той же генеральной совокупности.

В качестве статистики критерия используется величина

$$\chi^2 = n\sum_{i=1}^{m}\sum_{j=1}^{l}\frac{(n_{ij} - n_{i*}n_{*j}/n)^2}{n_{i*}n_{*j}} = n\left(\sum_{i=1}^{m}\sum_{j=1}^{l}\frac{n_{ij}^2}{n_{i*}n_{*j}} - 1\right), \quad \textbf{(10.22)}$$

где $n_{i*} = \sum_{j=1}^{l} n_{ij}$, $n_{*j} = \sum_{i=1}^{m} n_{ij}$, $n = \sum_{i=1}^{m} n_{i*} = \sum_{j=1}^{l} n_{*j}$.

В случае справедливости гипотезы H_0 статистика **(10.22)** имеет распределение χ^2 с $(m-1)(l-1)$ степенями свободы.

Пример 10.14а. По данным примера **10.14** на уровне значимости α=0,05 проверить гипотезу H_0 об однородности двух выборок (результатов двух проверок торговых точек города).

Р е ш е н и е. Необходимые для расчета статистики χ^2 величины представлены в табл. 10.8.

Таблица 10.8

Интервалы		0—10	10—20	20—30	30—40	40—50	50—60	60—70	70—80	80—90	$n_{*j} = \sum_{i=1}^{9} n_{ij}$
Частоты	n_{i1}	3	10	15	20	12	5	25	15	5	110
	n_{i2}	5	12	8	25	10	8	20	7	5	100
$n_{i*} = \sum_{i=1}^{2} n_{ij}$		8	22	23	45	22	13	45	22	10	n=210

По формуле **(10.22)** статистика критерия

$$\chi^2 = 210\left(\frac{3^2}{8\cdot 110} + \frac{10^2}{22\cdot 110} + \ldots + \frac{5^2}{10\cdot 110} + \frac{5^2}{8\cdot 100} + \right.$$

$$\left. + \frac{12^2}{22\cdot 100} + \ldots + \frac{5^2}{10\cdot 100} - 1\right) = 7{,}25.$$

По таблице V приложений при числе степеней свободы $(l-1)(m-1) = (9-1)(2-1) = 8$ $\chi^2_{0,05;8} = 15{,}5$. Так как $\chi^2 < \chi^2_{0,05;8}$, то гипотеза H_0 об однородности двух выборок не отвергается. ▶

Наряду с рассмотренными, в математической статистике используются также *ранговые*[1] критерии однородности, например, *критерии Вилкоксона—Манна—Уитни, Крускала—Уоллиса* и др. (см., например, [1]).

К ранговым относятся также ряд критериев **проверки гипотез о стохастической независимости элементов выборки**, таких как: *критерий серий, основанный на медиане выборки; критерий «восходящих» и «нисходящих» серий; критерий Аббе* (см. [1]). Рассмотрение вышеназванных критериев выходит за рамки данной книги.

В заключение отметим, что при проверке ряда гипотез, например, гипотез о законе распределения на заданном уровне значимости, контролируется лишь ошибка первого рода, но нельзя сделать вывод о степени риска, связанного с принятием неверной альтернативной гипотезы, т.е. с возможностью совершения ошибки второго рода.

Упражнения

10.15. По выборкам объемом n_1=14 и n_2=9 найдены средние размеры деталей соответственно $\bar{x}$=182 и y=185 мм, изготовленных на первом и втором автоматах. Установлено, что размер детали, изготовленной каждым автоматом, имеет нормальный закон распределения. Известны дисперсии $\sigma_x^2 = 5$ и $\sigma_y^2 = 7$ для первого и второго автоматов. На уровне значимости 0,05 выявить влияние на средний размер детали автомата, на котором она изготовлена. Рассмотреть два случая: а) конкурирующая гипотеза H_1: $x_0 \neq y_0$; б) конкурирующая гипотеза H_1: $x_0 < y_0$.

10.16. Расход сырья на единицу продукции составил:

по старой технологии

x_i	303	307	308	Всего
n_i	1	4	4	9

по новой технологии

y_j	303	304	306	308	Всего
n_j	2	6	4	1	13

[1] Ранговый критерий основан не на значениях признака, полученных в выборке, а на порядковых номерах (рангах) этих значений, расположенных в порядке возрастания (или убывания). Примеры использования рангов (правда, для других целей) приводятся в § 12.8.

Полагая, что расходы сырья по каждой технологии имеют нормальные распределения с одинаковыми дисперсиями, на уровне значимости 0,05 выяснить, дает ли новая технология экономию в среднем расходе сырья.

10.17. В рекламе утверждается, что месячный доход по акциям A превышает доход по акциям B более чем на 0,3% (или на 0,003). В течение годичного периода средний месячный доход по акциям B составил 0,5%, а по акциям A — 0,65%, а его средние квадратические отклонения соответственно 1,9 и 2,0%. Полагая распределения доходности по каждой акции нормальными, на уровне значимости 0,05 проверить утверждение, содержащееся в рекламе.

10.18. Имеются следующие данные о качестве детского питания, изготовленного различными фирмами (в баллах): 40, 39, 42, 37, 38, 43, 45, 41, 48. Есть основание полагать, что показатель качества продукции последней фирмы (48) зарегистрирован неверно. Является ли это значение аномальным (резко выделяющимся) на 5%-ном уровне значимости?

10.19. Вступительный экзамен проводился на двух факультетах института. На финансово-кредитном факультете из n_1=900 абитуриентов выдержали экзамен m_1=500 человек; а на учетно-статистическом факультете из n_2=800 абитуриентов — m_2=408. На уровне значимости $\alpha = 0{,}05$ проверить гипотезу об отсутствии существенных различий в уровне подготовки абитуриентов двух факультетов. Рассмотреть два случая: а) конкурирующая гипотеза H_1: $p_1 \neq p_2$; б) конкурирующая гипотеза H_1: $p_1 > p_2$.

10.20. В результате выборочной проверки качества однотипных изделий оказалось, что из 300 изделий фирмы A бракованных 30, из 400 фирмы B — 52, из 250 фирмы C — 21 и из 500 изделий фирмы D бракованных 74 изделия. На уровне значимости 0,05 выяснить, можно ли считать, что различия в качестве изделий различных фирм существенны.

10.21. По данным примера **10.16** выяснить, являются ли существенными различия между дисперсиями расхода

сырья на единицу продукции при использовании старой и новой технологий: а) на уровне значимости 0,05 при конкурирующей гипотезе $\sigma_x^2 > \sigma_y^2$; б) на уровне значимости 0,02 при конкурирующей гипотезе $\sigma_x^2 \neq \sigma_y^2$.

10.22. Сравниваются четыре способа обработки изделий. Лучшим считается тот из способов, при котором дисперсия контролируемого параметра меньше. Первым способом обработано 15 изделий, вторым — 20, третьим — 20, четвертым способом — 14 изделий. Выборочные дисперсии контролируемого параметра при разных способах обработки соответственно равны 26, 39, 48, 31 единиц. На уровне значимости 0,05 выяснить, можно ли считать, что способы обработки деталей обладают существенно различными дисперсиями. Можно ли признать первый способ «лучшим»? Предполагается, что контролируемый параметр распределен нормально.

10.23. Установлено, что средний вес таблетки лекарства сильного действия (номинал) должен быть равен 0,5 мг. Выборочная проверка n = 100 таблеток показала, что средний вес таблетки $x = 0{,}53$ мг. На основе проведенных исследований можно считать, что вес таблетки есть нормально распределенная случайная величина со средним квадратическим отклонением $\sigma_x = 0{,}11$ мг. На уровне значимости 0,05: а) выяснить, можно ли считать полученное в выборке отклонение от номинала случайным; б) найти мощность критерия, использованного в п. а).

10.24. Решить задачу **10.23** при условии, что n=20, x = 0,53 мг, а выборочное среднее квадратическое отклонение s_x = 0,11 мг.

10.25. Компания не осуществляет инвестиционных вложений в ценные бумаги с дисперсией годовой доходности более чем 0,04. Выборка из 52 наблюдений по активу A показала, что выборочная дисперсия ее доходности равна 0,045. Выяснить, допустимы ли для данной компании инвестиционные вложения в актив A на уровне значимости: а) 0,05; б) 0,01.

10.26. Фирма рассылает рекламные каталоги возможным заказчикам. Как показал опыт, вероятность того, что организация, получившая каталог, закажет рекламируемое изделие, равна 0,08. Фирма разослала 1000 каталогов новой, улучшенной, формы и получила 100 заказов. На уровне значимости 0,05 выяснить, можно ли считать, что новая форма рекламы существенно лучше прежней.

10.27. В соответствии со стандартом содержание активного вещества в продукции должно составлять 10%. Выборочная контрольная проверка 100 проб показала содержание активного вещества 15%. На уровне значимости 0,05 выяснить, должна ли продукция быть забракована. Рассмотреть два случая: а) конкурирующая гипотеза $p_1 \neq 0{,}1$; б) конкурирующая гипотеза $p_1 > 0{,}1$.

В примерах **10.28—10.30** на уровне значимости 0,05 проверить гипотезу о нормальном законе распределения признака (случайной величины) X, используя критерий согласия: а) χ^2-Пирсона; б) Колмогорова:

10.28. По данным примера **8.11**.

10.29. По данным примера **8.12**.

10.30. По данным примера **9.30**.

10.31. По данным примера **9.34** на уровне значимости 0,05 проверить гипотезу о показательном законе распределения признака (случайной величины) X, используя критерий: а) χ^2-Пирсона; б) Колмогорова.

10.32. Имеются две выборки значений (в усл. ед.) объемов 125 и 80 показателя качества однотипной продукции, изготовленной двумя фирмами:

x_i	14	17	20	23	26	29	32	35	38	41
n_i	2	4	10	15	20	27	18	16	8	5

y_j	16	20	24	28	32	36	40	44
n_j	3	9	12	17	16	13	7	3

Выяснить, можно ли на уровне значимости 0,05 считать, что рассматриваемый показатель качества продукции двух фирм описывается одной и той же функ-

цией распределения (т.е. выборки извлечены из одной генеральной совокупности). Решить задачу, используя критерии: а) Колмогорова—Смирнова; б) однородности χ^2.

10.33. Имеются следующие данные о числе сданных экзаменов в сессию студентами-заочниками:

Число сданных экзаменов x_i	0	1	2	3	4	Σ
Число студентов n_i	1	1	1	3	35	60

На уровне значимости $\alpha = 0{,}05$ проверить гипотезу о том, что случайная величина X — число сданных студентами экзаменов — распределена по биномиальному закону, используя критерий: а) χ^2-Пирсона; б) Колмогорова.

10.34. Имеются следующие данные о засоренности партии семян клевера семенами сорняков:

Число семян в одной пробе x_i	0	1	2	3	4	5	6	Σ
Число проб n_i	405	366	175	40	8	4	2	1000

На уровне значимости $\alpha = 0{,}05$ проверить гипотезу о том, что случайная величина X — число семян сорняков — распределена по закону Пуассона, используя критерий: а) χ^2-Пирсона; б) Колмогорова.

10.35. Фирма-производитель утверждает, что среднее время безотказной работы производимых ею электробытовых приборов составляет по меньшей мере 800 ч со средним квадратическим отклонением $\sigma = 120$ ч. Для случайно отобранных $n = 50$ приборов выборочное среднее время безотказной работы приборов оказалось равным 750 ч. На уровне значимости $\alpha = 0{,}05$: а) выяснить, удовлетворяет ли гарантии вся партия электробытовых приборов; б) найти мощность критерия, использованного в п. а); в) определить минимальное число приборов, которое следует проверить, чтобы обеспечить мощность критерия, равную 0,98.

10.36. Решить задачу **10.35** при $n = 15$, если σ неизвестно, а $s = 110$ получено по данным выборки.

Глава 11 Дисперсионный анализ

Выше рассмотрена проверка значимости (существенности, достоверности) различия выборочных средних двух совокупностей. На практике часто возникает необходимость обобщения задачи, т.е. проверки существенности различия выборочных средних m совокупностей (m>2). Например, требуется оценить влияние различных плавок на механические свойства металла, свойств сырья на показатели качества продукции, количества вносимых удобрений на урожайность и т.п.

Для эффективного решения такой задачи нужен новый подход, который и реализуется в дисперсионном анализе.

В настоящее время ***дисперсионный анализ*** *определяется как статистический метод, предназначенный для оценки влияния различных факторов на результат эксперимента, а также для последующего планирования аналогичных экспериментов.*

Первоначально (1918 г.) дисперсионный анализ был разработан английским математиком — статистиком Р.А. Фишером для обработки результатов агрономических опытов по выявлению условий получения максимального урожая различных сортов сельскохозяйственных культур. Сам термин «дисперсионный анализ» Фишер употребил позднее.

По числу факторов, влияние которых исследуется, различают однофакторный и многофакторный дисперсионный анализ.

11.1. Однофакторный дисперсионный анализ

Однофакторная дисперсионная модель имеет вид:

$$x_{ij} = \mu + F_i + \varepsilon_{ij}, \qquad \textbf{(11.1)}$$

где x_{ij} — значение исследуемой переменной, полученной на i-м уровне фактора (i=1,2,...,m) с j-м порядковым номером (j=1,2,...,n);

F_i — эффект, обусловленный влиянием i-го уровня фактора;

ε_{ij} — случайная компонента, или возмущение, вызванное влиянием неконтролируемых факторов, т.е. вариацией переменной внутри отдельного уровня.

Под *уровнем фактора* понимается некоторая его мера или состояние, например, количество вносимых удобрений, вид плавки металла или номер партии деталей и т.п.

Основные предпосылки дисперсионного анализа:

1. *Математическое ожидание возмущения* ε_{ij} *равно нулю для любых i, т.е.*

$$M(\varepsilon_{ij}) = 0. \qquad \textbf{(11.2)}$$

2. *Возмущения* ε_{ij} *взаимно независимы.*

3. *Дисперсия возмущения* ε_{ij} (*или переменной* x_{ij}) *постоянна для любых i, j, т.е.*

$$D(\varepsilon_{ij}) = \sigma^2. \qquad \textbf{(11.3)}$$

4. *Возмущение* ε_{ij} (*или переменная* x_{ij}) *имеет нормальный закон распределения* $N(0;\sigma^2)$.

Влияние уровней фактора может быть как *фиксированным*, или *систематическим* (модель I), так и *случайным* (модель II).

Пусть, например, необходимо выяснить, имеются ли существенные различия между партиями изделий по некоторому показателю качества, т.е. проверить влияние на качество одного фактора — партии изделий. Если включить в исследование все партии сырья, то влияние уровня такого фактора систематическое (модель I), а полученные выводы применимы только к тем отдельным партиям, которые привлекались при исследовании; если же включить только отобранную случайно часть партий, то влияние фактора случайное (модель II). В многофакторных комплексах возможна смешанная модель III, в которой одни факторы имеют случайные уровни, а другие — фиксированные.

Рассмотрим эту задачу подробнее. Пусть имеется m партий изделий. Из каждой партии отобрано соответственно $n_1,n_2,...,n_m$ изделий (для простоты полагаем, что $n_1=n_2=...=n_m=n$). Значения показателя качества этих изделий представим в виде матрицы наблюдений

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix} = (x_{ij}), \ (i = 1,2,...,m; \ j = 1,2,...,n).$$

Необходимо проверить существенность влияния партий изделий на их качество.

Если полагать, что элементы строк матрицы наблюдений — это численные значения (реализации) случайных величин $X_1, X_2, ..., X_m$, выражающих качество изделий и имеющих нормальный закон распределения с математическими ожиданиями соответственно $a_1, a_2, ..., a_m$ и одинаковыми дисперсиями σ^2, то данная задача сводится к проверке нулевой гипотезы H_0: $a_1 = a_2 = ... = a_m$, осуществляемой в дисперсионном анализе.

Обозначим усреднение по какому-либо индексу звездочкой (или точкой) вместо индекса, тогда средний показатель качества изделий i-й партии, или групповая средняя для i-го уровня фактора, примет вид:

$$\bar{x}_{i*} = \frac{\sum_{j=1}^{n} x_{ij}}{n}, \qquad (11.4)$$

а общая средняя —

$$\bar{x}_{**} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}}{mn} = \frac{\sum_{i=1}^{m} \bar{x}_{i*}}{m}. \qquad (11.5)$$

Рассмотрим сумму квадратов отклонений наблюдений x_{ij} от общей средней $\bar{x}_{**}$:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{**})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{x}_{i*} - \bar{x}_{**})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*})^2 + 2\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{**})(\bar{x}_{i*} - \bar{x}_{**})^2, \qquad (11.6)$$

или $Q = Q_1 + Q_2 + Q_3$.

Последнее слагаемое

$$Q_3 = 2\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*})(\bar{x}_{i*} - \bar{x}_{**}) = 2\sum_{i=1}^{m}(\bar{x}_{i*} - \bar{x}_{**})\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*}) = 0,$$

так как сумма отклонений значений переменной от ее средней, т.е. $\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*})$ равна нулю.

Первое слагаемое можно записать в виде:

$$Q_1 = \sum_{i=1}^{m}\sum_{j=1}^{n}(\bar{x}_{i*} - \bar{x}_{**})^2 = n\sum_{i=1}^{m}(\bar{x}_{i*} - \bar{x}_{**})^2. \qquad (11.7)$$

В результате получим следующее тождество:

$$Q = Q_1 + Q_2, \qquad (11.8)$$

где $Q = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{**})^2$ — *общая*, или *полная*, сумма квадратов отклонений;

$Q_1 = n\sum_{i=1}^{m}(\bar{x}_{i*} - \bar{x}_{**})^2$ — сумма квадратов отклонений групповых средних от общей средней, или *межгрупповая* (*факторная*) сумма квадратов отклонений;

$Q_2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i*})^2$ — сумма квадратов отклонений наблюдений от групповых средних, или *внутригрупповая* (*остаточная*) сумма квадратов отклонений.

В разложении **(11.8)** заключена основная идея дисперсионного анализа. Если поделить обе части равенства **(11.8)** на число наблюдений, то получим рассмотренное выше правило сложения дисперсий **(8.12)**. Применительно к рассматриваемой задаче равенство **(11.8)** показывает, что общая вариация показателя качества, измеренная суммой Q, складывается из двух компонент — Q_1 и Q_2, характеризующих изменчивость этого показателя между партиями (Q_1) и изменчивость «внутри» партий (Q_2), характеризующих одинаковую (по условию) для всех партий вариацию под воздействием неучтенных факторов.

В дисперсионном анализе анализируются не сами суммы квадратов отклонений, а так называемые средние квадраты, являющиеся несмещенными оценками соответствующих дисперсий, которые получаются делением сумм квадратов отклонений на соответствующее число степеней свободы.

Напомним, что число степеней свободы определяется как общее число наблюдений минус число связывающих их уравнений. Поэтому для среднего квадрата s_1^2, являющегося несмещенной оценкой межгрупповой дисперсии, число степеней свободы $k_1 = m - 1$, так как при его расчете используются m групповых средних, связанных между собой одним уравнением **(11.5)**. А для среднего квадрата s_2^2, являющегося несмещенной оцен-

кой внутригрупповой дисперсии, число степеней свободы $k_2=mn-m$, ибо при ее расчете используются все mn наблюдений, связанных между собой m уравнениями **(11.4)**. Таким образом, $s_1^2=Q_1/(m-1)$; $s_2^2=Q_2/(mn-m)$.

Найдем математические ожидания средних квадратов s_1^2 и s_2^2, подставив в их формулы выражение x_{ij} **(11.1)** через параметры модели.

$$M\left(s_1^2\right)=\frac{n}{m-1}M\left(\sum_{i=1}^{m}\left(\mu+F_i+\varepsilon_{i*}-\mu-F_*-\varepsilon_{**}\right)^2\right)=$$

$$=\frac{n}{m-1}M\left[\sum_{i=1}^{m}\left(\left(F_i-F_*\right)+\left(\varepsilon_{i*}-\varepsilon_{**}\right)\right)^2\right]=\frac{n}{m-1}M\left[\sum_{i=1}^{m}\left(F_i-F_*\right)^2\right]+$$

$$+\frac{n}{m-1}M\left[2\sum_{i=1}^{m}\left(F_i-F_*\right)\left(\varepsilon_{i*}-\varepsilon_{**}\right)\right]+\frac{n}{m-1}M\left[\sum_{i=1}^{m}\left(\varepsilon_{i*}-\varepsilon_{**}\right)^2\right]=$$

$$=\frac{n}{m-1}M\left[\sum_{i=1}^{m}\left(F_i-F_*\right)^2\right]+\sigma^2 \qquad \textbf{(11.9)}$$

(ибо $M\left[2\sum_{i=1}^{m}\left(F_i-F_*\right)\left(\varepsilon_{i*}-\varepsilon_{**}\right)\right]=0$ с учетом свойств математического ожидания, а

$$\frac{n}{m-1}M\left[\sum_{i=1}^{m}\left(\varepsilon_{i*}-\varepsilon_{**}\right)^2\right]=n\cdot M\left[\frac{\sum_{i=1}^{m}\left(\varepsilon_{i*}-\varepsilon_{**}\right)^2}{m-1}\right]=n\sigma^2_{\varepsilon_{i*}}=n\frac{\sigma^2_{\varepsilon_{ij}}}{n}=\sigma^2).$$

$$M\left(s_2^2\right)=\frac{1}{m(n-1)}M\left[\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\mu+F_i+\varepsilon_{ij}-\mu-F_i-\varepsilon_{i*}\right)^2\right]=$$

$$=\frac{1}{m}M\left[\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\varepsilon_{ij}-\varepsilon_{**}\right)^2}{n-1}\right]=\frac{1}{m}\sum_{i=1}^{m}\sigma^2_{\varepsilon_{ij}}=\frac{1}{m}\sum_{i=1}^{m}\sigma^2=\frac{1}{m}\cdot m\sigma^2=\sigma^2 \quad . \qquad \textbf{(11.10)}$$

Схему дисперсионного анализа представим в виде таблицы.

Таблица 11.1

Компоненты дисперсии	*Сумма квадратов*	*Число степеней свободы*	*Средний квадрат*	*Математическое ожидание среднего квадрата*
Межгрупповая	$Q_1=n\sum_{i=1}^{m}\left(\bar{x}_{i*}-\bar{x}_{**}\right)^2$	$m-1$	$s_1^2=\frac{Q_1}{m-1}$	$M\left(s_1^2\right)=\begin{cases}\frac{n}{m-1}\sum_{i=1}^{m}\left(F_i-F_*\right)^2+\sigma^2 & \text{(модель I)},\\ n\sigma_F^2+\sigma^2 & \text{(модель II)}\end{cases}$
Внутригрупповая	$Q_2=\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_{i*}\right)^2$	$mn-m$	$s_2^2=\frac{Q_2}{mn-m}$	$M\left(s_2^2\right)=\sigma^2$
Общая	$Q=\sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_{**}\right)^2$	$mn-1$		

Для модели I с фиксированными уровнями фактора F_i ($i=1,2,...,m$) — величины неслучайные, поэтому

$$M\left(s_1^2\right)=n\sum_{i=1}^{n}\left(F_i-F_*\right)^2/(m-1)+\sigma^2 .$$

Гипотеза H_0 примет вид $F_i=F_*\left(i=1,2,...,m\right)$, т.е. влияние всех уровней фактора одно и то же. В случае справедливости этой гипотезы $M\left(s_1^2\right)=M\left(s_2^2\right)=\sigma^2$.

Для случайной модели II слагаемое F_i в выражении **(11.1)** — величина случайная. Обозначая ее дисперсию

$\sigma_F^2=M\left[\sum_{i=1}^{m}\left(F_i-F_*\right)^2/(m-1)\right]$, получим из **(11.9)**

$$M\left(s_1^2\right)=n\sigma_F^2+\sigma^2, \qquad \textbf{(11.11)}$$

и, как и в модели I, $M\left(s_2^2\right)=\sigma^2$. В случае справедливости нулевой гипотезы H_0, которая для модели II принимает вид $\sigma_F^2=0$, имеем: $M\left(s_1^2\right)=M\left(s_2^2\right)=\sigma^2$.

Итак, в случае однофакторного комплекса как для модели I, так и модели II *средние квадраты* s_1^2 и s_2^2 *являются несмещенными* и, как можно показать, *независимыми оценками одной и той же дисперсии* σ^2.

Следовательно, проверка нулевой гипотезы H_0 свелась к проверке существенности различия несмещенных выборочных оценок s_1^2 и s_2^2 дисперсии σ^2, рассмотренной в § 10.5.

Гипотеза H_0 отвергается, если фактически вычисленное значение статистики $F=\frac{s_1^2}{s_2^2}$ больше критического $F_{\alpha;k_1;k_2}$, определенного на уровне значимости α при числе степеней свободы $k_2=mn-m$, и принимается, если $F \le F_{\alpha;k_1;k_2}$.

Применительно к данной задаче опровержение гипотезы H_0 означает наличие существенных различий в качестве изделий различных партий на рассматриваемом уровне значимости.

Замечание. Для вычисления сумм квадратов Q_1, Q_2, Q часто бывает удобно использовать следующие формулы:

$$Q_1=\frac{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}x_{ij}\right)^2}{n}-\frac{\left(\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}\right)^2}{mn}, \qquad \textbf{(11.12)}$$

$$Q_2=\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}{}^2-\frac{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}x_{ij}\right)^2}{n}, \qquad \textbf{(11.13)}$$

$$Q=\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}^2-\frac{\left(\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}\right)^2}{mn}, \qquad \textbf{(11.14)}$$

т.е. сами средние, вообще говоря, находить не обязательно.

▷ **Пример 11.1.** Имеются 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. 11.2.

Таблица 11.2

Номер партии	*Разрывная нагрузка* (кг/см^2)				
1	200	140	170	145	165
2	190	150	210	150	150
3	230	190	200	190	200
4	150	170	150	170	180

Необходимо выяснить, существенно ли влияние различных партий сырья на величину разрывной нагрузки. Принять $\alpha = 0{,}05$.

Решение. Имеем $m=4$, $n=5$. Найдем средние значения разрывной нагрузки для каждой партии по формуле **(11.4)**:

$$\bar{x}_{1*}=(200+140+170+145+165)/5 = 164 \text{ (кг/см}^2\text{)}$$

и аналогично

$$\bar{x}_{2*}= 170,\quad \bar{x}_{3*}=202 \text{ и } \bar{x}_{4*}= 164 \text{ (кг/см}^2\text{)}.$$

Среднее значение разрывной нагрузки всех отобранных образцов по формуле **(11.5)**:

$$\bar{x}_{**}=(200+140+\ldots+170+180)/20=175 \text{ (кг/см}^2\text{)}$$

(или, иначе, через групповые средние,

$$x_{**}=(164+170+202+164)/4=175 \text{ (кг/см}^2\text{)}).$$

Вычислим суммы квадратов отклонений по формулам **(11.6)**, **(11.7)**:

$$Q_1=5\sum_{i=1}^{4}(\bar{x}_{i*}-\bar{x}_{**})^2=5\left[(164-175)^2+(170-175)^2+(202-175)^2+\right.$$

$$\left.+(164-175)^2\right]=5\cdot 996=4980;$$

$$Q_2=\sum_{i=1}^{4}\sum_{j=1}^{5}(x_{ij}-\bar{x}_{i*})^2=(200-164)^2+\ldots+(165-164)^2+$$

$$+(190-170)^2+\ldots+(150-170)^2+(230-202)^2+\ldots+(200-202)^2+$$

$$+(150-164)^2+\ldots+(180-164)^2=7270;$$

$$Q=\sum_{i=1}^{4}\sum_{j=1}^{5}(x_{ij}-\bar{x}_{**})^2=(200-175)^2+(140-175)^2+\ldots+(170-175)^2+$$

$$+(180-175)^2=12\,250.$$

Соответствующее число степеней свободы для этих сумм $m-1=3$; $mn-m=5\cdot 4-4=16$; $mn-1=5\cdot 4-1=19$.

Результаты расчета сведем в табл. 11.3.

Таблица 11.3

Компоненты дисперсии	*Суммы квадратов*	*Число степеней свободы*	*Средние квадраты*
Межгрупповая	4980	3	1660,0
Внутригрупповая	7270	16	454,4
Общая	12250	19	

Фактически наблюдаемое значение статистики $F=\frac{s_1^2}{s_2^2}=\frac{1660}{454{,}4}=$ $=3{,}65$. По табл. VI приложений критическое значение F-критерия

Фишера—Снедекора на уровне значимости $\alpha=0{,}05$ при $k_1=3$ и $k_2=16$ степенях свободы $F_{0,05;3;16}=3{,}24$. Так как $F > F_{0,05;3;16}$, то нулевая гипотеза отвергается, т.е. на уровне значимости $\alpha=0{,}05$ (с надежностью 0,95) различие между партиями сырья оказывает существенное влияние на величину разрывной нагрузки.

Замечание. С точки зрения техники вычислений сумм Q_1, Q_2, Q проще воспользоваться формулами **(11.12)—(11.14)**, не требующими вычисления средних. Так, вычислив

$$\sum_{i=1}^{4}\sum_{j=1}^{5} x_{ij} = 200+140+\ldots+170+180=3500,$$

$$\sum_{i=1}^{4}\sum_{j=1}^{5} x_{ij}^2 = 200^2+140^2+\ldots+170^2+180^2=624\,750,$$

$$\sum_{i=1}^{4}\left(\sum_{j=1}^{5} x_{ij}\right)^2 = (200+\ldots+165)^2+(190+\ldots+150)^2+$$

$$+(230+\ldots+200)^2+(150+\ldots+180)^2=3\,087\,400,$$

найдем

по **(11.12)** $Q_1=3\,087\,400/5-3500^2/20=4980$,

по **(11.13)** $Q_2=624\,750-3\,087\,400/5=7270$

и по **(11.14)** $Q=624\,750-3500^2/20=12\,250$. ▶

11.2. Понятие о двухфакторном дисперсионном анализе

Предположим, что в рассматриваемой в § 11.1 задаче о качестве различных (m) партий изделия изготавливались на разных (l) станках и требуется выяснить, имеются ли существенные различия в качестве изделий по каждому фактору: A — партия изделий, B — станок. В результате мы приходим к задаче *двухфакторного дисперсионного анализа.*

Все имеющиеся данные представим в виде табл. 11.4, в которой по строкам — уровни A_i фактора A, по столбцам — уровни B_j фактора B, а в соответствующих клетках, или ячейках, таблицы находятся значения показателя качества изделий x_{ijk} ($i=1,2,\ldots,m$; $j=1,2,\ldots,l$; $k=1,2,\ldots,n$):

Таблица 11.4

A \ B	B_1	B_2	...	B_j	...	B_l
A_1	$x_{111},\ldots,x_{11k}$	$x_{121},\ldots,x_{12k}$	...	$x_{1j1},\ldots,x_{1jk}$	...	$x_{1l1},\ldots,x_{1lk}$
A_2	$x_{211},\ldots,x_{21k}$	$x_{221},\ldots,x_{22k}$	...	$x_{2j1},\ldots,x_{2jk}$	...	$x_{2l1},\ldots,x_{2lk}$
⋮	...	...	...	...	...	...
A_i	$x_{i11},\ldots,x_{i1k}$	$x_{i21},\ldots,x_{i2k}$	...	$x_{ij1},\ldots,x_{ijk}$	...	$x_{il1},\ldots,x_{ilk}$
⋮	...	...	...		...	...
A_m	$x_{m11},\ldots,x_{m1k}$	$x_{m21},\ldots,x_{m2k}$	...	$x_{mj1},\ldots,x_{mjk}$	...	$x_{ml1},\ldots,x_{mlk}$

Двухфакторная дисперсионная модель имеет вид:

$$x_{ijk}=\mu+F_i+G_j+I_{ij}+\varepsilon_{ijk}, \qquad \textbf{(11.15)}$$

где x_{ijk} — значение наблюдения в ячейке ij с номером k;

μ — общая средняя;

F_i — эффект, обусловленный влиянием i-го уровня фактора A;

G_j — эффект, обусловленный влиянием j-го уровня фактора B;

I_{ij} — эффект, обусловленный взаимодействием двух факторов, т.е. отклонение от средней по наблюдениям в ячейке ij от суммы первых трех слагаемых в модели **(11.15)**;

ε_{ijk} — возмущение, обусловленное вариацией переменной внутри отдельной ячейки.

Полагаем, что ε_{ijk} имеет нормальный закон распределения $N(0;\sigma^2)$, а все математические ожидания F_*, G_*, I_{i*}, I_{*j} равны нулю.

Групповые средние находятся по формулам:
в ячейке —

$$\bar{x}_{ij_*}=\frac{\sum_{k=1}^{n} x_{ijk}}{n}, \qquad \textbf{(11.16)}$$

по строке —

$$\bar{x}_{i**}=\frac{\sum_{j=1}^{l} \bar{x}_{ij*}}{l}, \qquad \textbf{(11.17)}$$

по столбцу —

$$\bar{x}_{*j*} = \frac{\sum_{i=1}^{m} \bar{x}_{ij*}}{m}. \qquad (11.18)$$

Общая средняя

$$\bar{x}_{***} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{l} \bar{x}_{ij*}}{ml}. \qquad (11.19)$$

Таблица дисперсионного анализа имеет вид:

Таблица 11.5

Компоненты дисперсии	*Сумма квадратов*	*Число степеней свободы*	*Средние квадраты*
Межгрупповая (*фактор A*)	$Q_1 = \ln\sum_{i=1}^{m}\left(\bar{x}_{i**} - \bar{x}_{***}\right)^2$	$m-1$	$s_1^2 = \frac{Q_1}{m-1}$
Межгрупповая (*фактор B*)	$Q_2 = mn\sum_{j=1}^{l}\left(\bar{x}_{*j*} - \bar{x}_{***}\right)^2$	$l-1$	$s_2^2 = \frac{Q_2}{l-1}$
Взаимодействие (*AB*)	$Q_3 = n\sum_{i=1}^{m}\sum_{j=1}^{l}\left(\bar{x}_{ij*} - \bar{x}_{i**} - \bar{x}_{*j*} + \bar{x}_{***}\right)^2$	$(m-1)(l-1)$	$s_3^2 = \frac{Q_3}{(m-1)(l-1)}$
Остаточная	$Q_4 = \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n}\left(x_{ijk} - \bar{x}_{ij*}\right)^2$	$mln-ml$	$s_4^2 = \frac{Q_4}{mln-ml}$
Общая	$Q = \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n}\left(x_{ijk} - \bar{x}_{***}\right)^2$	$mln-1$	

Можно показать, что проверка нулевых гипотез H_A, H_B, H_{AB} об отсутствии влияния на рассматриваемую переменную факторов A, B и их взаимодействия AB осуществляется сравнением отношений $s_1^2/s_4^2, s_2^2/s_4^2, s_3^2/s_4^2$ (для модели I с фиксированными уровнями факторов) или отношений $s_1^2/s_3^2, s_2^2/s_3^2, s_3^2/s_4^2$ (для случайной модели II) с соответствующими табличными значениями F-критерия Фишера—Снедекора. Для смешанной модели III проверка гипотез относительно факторов с фиксированными уровнями проводится

так, как в модели II, а факторов со случайными уровнями — как в модели I.

Если $n=1$, т.е. при одном наблюдении в ячейке, то не все нулевые гипотезы могут быть проверены, так как выпадает компонента Q_3 из общей суммы квадратов отклонений, а с ней и средний квадрат s_3^2, ибо в этом случае не может быть речи о взаимодействии факторов.

▷ **Пример 11.2.** В табл. 11.6 приведены суточные привесы (г) отобранных для исследования 18 поросят в зависимости от метода содержания поросят (фактор *A*) и качества их кормления (фактор *B*).

Таблица 11.6

Количество голов в группе (фактор A)	*Содержание протеина в корме, г (фактор B)*	
	B_1=80	B_2=100
A_1 = 30	530, 540, 550	600, 620, 580
A_2 = 100	490, 510, 520	550, 540, 560
A_3 = 300	430, 420, 450	470, 460, 430

Необходимо на уровне значимости $\alpha = 0{,}05$ оценить существенность (достоверность) влияния каждого фактора и их взаимодействия на суточный привес поросят.

Р е ш е н и е. Имеем m=3, l=2, n=3. Определим (в г) средние значения привеса:

в ячейках — по (**11.16**):

$$\bar{x}_{11*} = \frac{530+540+550}{3} = 540 \text{ и аналогично } \bar{x}_{12*} = 600;$$

$$\bar{x}_{21*} = 506{,}7;\ \bar{x}_{22*} = 550;\ \bar{x}_{31*} = 433{,}3;\ \bar{x}_{32*} = 453{,}3;$$

по строкам — по (**11.17**):

$$\bar{x}_{1**} = \frac{540+600}{2} = 570 \text{ и аналогично } \bar{x}_{2**} = 528{,}4;\ \bar{x}_{3**} = 443{,}2;$$

по столбцам — по (**11.18**):

$$\bar{x}_{*1*} = \frac{540+506{,}7+433{,}3}{3} = 493{,}3 \text{ и аналогично } \bar{x}_{*2*} = 534{,}4.$$

Общий средний привес — по (**11.19**):

$$\bar{x}_{***} = \frac{540+600+506{,}7+550+433{,}3+453{,}3}{6} = 513{,}9\ (\text{г}).$$

Все средние значения привеса (г) поместим в табл. 11.7.

Таблица 11.7

Количество голов в группе (фактор A)	*Содержание протеина в корме, г (фактор B)*		
	B_1=80	B_2=100	x_{i**}
A_1=30	$\bar{x}_{11*}=540{,}0$	$\bar{x}_{12*}=600{,}0$	$\bar{x}_{1**}=570{,}0$
A_2=100	$\bar{x}_{21*}=506{,}7$	$\bar{x}_{22*}=550{,}0$	$\bar{x}_{2**}=528{,}4$
A_3=300	$\bar{x}_{31*}=433{,}3$	$\bar{x}_{32*}=453{,}3$	$\bar{x}_{3**}=443{,}3$
$\bar{x}_{*j*}$	$\bar{x}_{*1*}=493{,}3$	$\bar{x}_{*2*}=534{,}4$	$\bar{x}_{***}=513{,}9$

Из табл. 11.7 следует, что с увеличением количества голов в группе средний суточный привес поросят в среднем уменьшается, а при увеличении содержания протеина в корме — в среднем увеличивается. Но является ли эта тенденция достоверной или объясняется случайными причинами? Для ответа на этот вопрос по формулам табл. 11.5 вычислим необходимые суммы квадратов отклонений:

$$Q_1=2\cdot 3[(570-513{,}9)^2+(528{,}4-513{,}9)^2+(443{,}2-513{,}9)^2]=50\,011{,}1;$$

$$Q_2=3\cdot 3[(493{,}3-513{,}9)^2+(534{,}4-513{,}9)^2]=7605{,}6;$$

$$Q_3=3[(540-570-493{,}3+513{,}9)^2+\ldots+$$

$$+(453{,}3-443{,}3-534{,}4+513{,}9)^2]=1211{,}1;$$

$$Q_4=(530-540)^2+\ldots+(550-540)^2+(600-600)^2+\ldots+$$

$$+(580-600)^2+\ldots+(470-453{,}3)^2+\ldots+(430-453{,}3)^2=3000{,}0;$$

$$Q=(530-513{,}9)^2+(540-513{,}9)^2+\ldots+(430-513{,}9)^2=61\,827{,}8.$$

Средние квадраты находим делением полученных сумм на соответствующее им число степеней свободы $m-1=2$, $l-1=1$; $(m-1)(l-1)=2$; $mln-ml=18-6=12$; $mln-1=18-1=17$.

Результаты расчета сведем в табл. 11.8.

Таблица 11.8

Компонента дисперсии	*Суммы квадратов*	*Число степеней свободы*	*Средние квадраты*
Межгрупповая (*фактор A*)	Q_1=50 011,1	2	$s_1^2=25\,005{,}5$
Межгрупповая (*фактор B*)	Q_2=7605,6	1	$s_2^2=7605{,}6$
Взаимодействие (*AB*)	Q_3=1211,1	2	$s_3^2=605{,}6$
Остаточная	Q_4=3000,0	12	$s_4^2=250{,}0$
Общая	Q=61 827,8	17	

Очевидно, данные факторы имеют фиксированные уровни, т.е. мы находимся в рамках модели I. Поэтому для проверки существенности влияния факторов *A*, *B* и их взаимодействия *AB* необходимо найти отношения:

$$F_A=\frac{s_1^2}{s_4^2}=\frac{25\,005{,}5}{250{,}0}=100{,}0;\quad F_B=\frac{s_2^2}{s_4^2}=\frac{7605{,}6}{250{,}0}=30{,}4;$$

$F_{AB}=\dfrac{s_3^2}{s_4^2}=\dfrac{605{,}6}{250{,}0}=2{,}42$ и сравнить их с табличными значениями (см. табл. VI приложений) соответственно $F_{0{,}05;2;12}$=3,88; $F_{0{,}05;1;12}$= =4,75; $F_{0{,}05;2;12}$=3,88. Так как $F_A>F_{0{,}05;2;12}$ и $F_B>F_{0{,}05;1;12}$, то влияние метода содержания поросят (фактора *A*) и качества их кормления (фактора *B*) является существенным. В силу того что $F_{AB}<F_{0{,}05;2;12}$, взаимодействие указанных факторов незначимо (на 5%-ном уровне). ▶

Замечание. С точки зрения техники вычислений для нахождения сумм квадратов Q_1, Q_2, Q_3, Q_4, Q целесообразнее использовать формулы:

$$Q_1=\frac{\sum_{i=1}^{m}\left(\sum_{j=1}^{l}\sum_{k=1}^{n}x_{ijk}\right)^2}{ln}-\frac{\left(\sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n}x_{ijk}\right)^2}{mln},\qquad\textbf{(11.20)}$$

$$Q_2=\frac{\sum_{j=1}^{l}\left(\sum_{i=1}^{m}\sum_{k=1}^{n}x_{ijk}\right)^2}{mn}-\frac{\left(\sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n}x_{ijk}\right)^2}{mln},\qquad\textbf{(11.21)}$$

$$Q_4 = \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n} x_{ijk}^2 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{l}\left(\sum_{k=1}^{n} x_{ijk}\right)^2}{n}, \qquad \textbf{(11.22)}$$

$$Q = \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n} x_{ijk}^2 - \frac{\left(\sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n} x_{ijk}\right)^2}{mln}, \qquad \textbf{(11.23)}$$

$$Q_3 = Q - Q_1 - Q_2 - Q_4. \qquad \textbf{(11.24)}$$

Так, в рассматриваемом примере **11.2**:

$$\sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{3} x_{ijk} = 530+540+...+460+430 = 9250,$$

$$\sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{3} x_{ijk}^2 = 530^2+540^2+...+460^2+430^2 = 4\ 815\ 300,$$

$$\sum_{i=1}^{3}\left(\sum_{j=1}^{2}\sum_{k=1}^{3} x_{ijk}\right)^2 = (530+540+...+620+580)^2+...+$$

$$+(430+420+...+460+430)^2 = 28\ 820\ 900,$$

$$\sum_{j=1}^{2}\left(\sum_{i=1}^{3}\sum_{k=1}^{3} x_{ijk}\right)^2 = (530+540+...+420+450)^2+...+$$

$$+(600+620+...+460+430)^2 = 42\ 849\ 700,$$

$$\sum_{i=1}^{3}\sum_{j=1}^{2}\left(\sum_{k=1}^{3} x_{ijk}\right)^2 = (530+540+550)^2+(600+620+580)^2+...+$$

$$+(470+460+430)^2 = 14\ 436\ 900,$$

и по формулам **(11.20)**—**(11.24)**:

$$Q_1 = \frac{28\ 820\ 900}{2\cdot 3} - \frac{9250^2}{3\cdot 2\cdot 3} = 50\ 011{,}1;$$

$$Q_2 = \frac{42\ 849\ 700}{3\cdot 3} - \frac{9250^2}{3\cdot 2\cdot 3} = 7605{,}6;$$

$$Q_4 = 4\ 815\ 300 - \frac{14\ 436\ 900}{3} = 3000;\ Q = 4\ 815\ 300 - \frac{9250^2}{3\cdot 2\cdot 3} = 61\ 827{,}8;$$

$$Q_3 = 61\ 827{,}8 - 50\ 011{,}1 - 7605{,}6 - 3000 = 1211{,}1. \blacktriangleright$$

В заключение отметим, что при решении реальных задач методом дисперсионного анализа используются *статистические программные пакеты* (см., например, [30]).

Отклонение от основных предпосылок дисперсионного анализа — нормальности распределения исследуемой переменной и равенства дисперсий в ячейках (если оно не чрезмерное) — не сказывается существенно на результатах дисперсионного анализа при равном числе наблюдений в ячейках, но может быть очень чувствительно при неравном их числе. Кроме того, при неравном числе наблюдений в ячейках резко возрастает сложность аппарата дисперсионного анализа. Поэтому рекомендуется планировать схему с равным числом наблюдений в ячейках, а если встречаются недостающие данные, то возмещать их средними значениями других наблюдений в ячейках. При этом, однако, искусственно введенные недостающие данные не следует учитывать при подсчете числа степеней свободы.

Упражнения

11.3. В течение шести лет использовались пять различных технологий по выращиванию сельскохозяйственной культуры. Данные по эксперименту (в ц/га) приведены в таблице:

Номер наблюдения (год)	*Технология (фактор A)*				
	A_1	A_2	A_3	A_4	A_5
1	1,2	0,6	0,9	1,7	1,0
2	1,1	1,1	0,6	1,4	1,4
3	1,0	0,8	0,8	1,3	1,1
4	1,3	0,7	1,0	1,5	0,9
5	1,1	0,7	1,0	1,2	1,2
6	0,8	0,9	1,1	1,3	1,5
Итого	6,5	4,8	5,4	8,4	7,1

Необходимо на уровне значимости $\alpha = 0{,}05$ установить влияние различных технологий на урожайность культуры.

11.4. На заводе установлено четыре линии по выпуску облицовочной плитки. С каждой линии случайным образом в течение смены отобрано по 10 плиток и сделаны замеры их толщины (мм). Отклонения от номинального размера приведены в таблице:

Линия по выпуску плиток	Номер испытания									
	1	2	3	4	5	6	7	8	9	10
1	0,6	0,2	0,4	0,5	0,8	0,2	0,1	0,6	0,8	0,8
2	0,2	0,2	0,4	0,3	0,3	0,6	0,8	0,2	0,5	0,5
3	0,8	0,6	0,2	0,4	0,9	1,1	0,8	0,2	0,4	0,8
4	0,7	0,7	0,3	0,3	0,2	0,8	0,6	0,4	0,2	0,6

Требуется на уровне значимости $\alpha = 0{,}05$ установить зависимость выпуска качественных плиток от линии выпуска (фактора A).

11.5. Имеются следующие данные об урожайности четырех сортов пшеницы на выделенных пяти участках земли (блоках):

Сорт	Урожайность по блокам, ц/га				
	1	2	3	4	5
1	2,87	2,67	2,16	2,50	2,82
2	2,45	2,85	2,77	2,87	3,25
3	2,32	2,47	2,00	2,40	2,40
4	2,90	2,87	2,25	2,80	2,70

Требуется на уровне значимости $\alpha = 0{,}05$ установить влияние на урожайность сорта пшеницы (фактора A) и участков земли — блоков (фактора B).

11.6. На четырех предприятиях B_1, B_2, B_3, B_4 проверялись три технологии производства A_1, A_2, A_3 однотипных изделий. Данные о производительности труда в условных единицах приведены в таблице:

A / B	A_1			A_2			A_3		
	1	2	3	1	2	3	1	2	3
B_1	50	54	58	62	60	58	65	71	65
B_2	54	46	50	64	59	60	59	54	61
B_3	52	48	50	70	62	60	59	66	64
B_4	60	55	56	58	54	50	71	74	62

Требуется на уровне значимости $\alpha = 0{,}05$ установить влияние на производительность труда технологий (фактора A) и предприятий (фактора B).

Глава 12 Корреляционный анализ

Диалектический подход к изучению природы и общества требует рассмотрения явлений в их взаимосвязи и непрестанном изменении.

Понятия *корреляции* и *регрессии* появились в середине XIX в. благодаря работам английских статистиков Ф. Гальтона и К. Пирсона. Первый термин произошел от латинского «*correlatio*» — соотношение, взаимосвязь. Второй термин (от лат. «*regressio*» — движение назад) введен Ф. Гальтоном, который, изучая зависимость между ростом родителей и их детей, обнаружил явление «регрессии к среднему» — у детей, родившихся у очень высоких родителей, рост имел тенденцию быть ближе к средней величине.

12.1. Функциональная, статистическая и корреляционная зависимости

В естественных науках часто речь идет о ***функциональной*** зависимости (связи), когда каждому значению одной переменной соответствует вполне *определенное значение другой.* Функциональная зависимость может иметь место как между детерминированными (неслучайными) переменными (например, зависимость скорости падения в вакууме от времени и т.п.), так и между случайными величинами (например, зависимость стоимости проданных изделий от их числа и т.п.).

В экономике в большинстве случаев между переменными величинами существуют зависимости, когда каждому значению одной переменной соответствует не какое-то определенное, а м н о ж е с т в о возможных значений другой переменной. Иначе говоря, каждому значению одной переменной соответствует *определенное* (*условное*) *распределение другой переменной.* Такая зависимость (связь) получила название ***статистической*** (или ***стохастической, вероятностной***). (О ней уже шла речь в § 5.5.)

Возникновение понятия статистической связи обусловливается тем, что зависимая переменная подвержена влиянию ряда неконтролируемых или неучтенных факторов, а также тем, что измерение значений переменных неизбежно сопровождается некоторыми случайными ошибками. Примером статистической связи является зависимость урожайности от количества внесенных удобрений, производительности труда на предприятии от его энерговооруженности и т.п.

В силу неоднозначности статистической зависимости между Y и X для исследователя, в частности, представляет интерес усредненная по x схема зависимости, т.е. закономерность в изменении среднего значения — условного математического ожидания[1] $M_x(Y)$ (математического ожидания случайной переменной Y, вычисленного в предположении, что переменная X приняла значение x) в зависимости от x.

Определение. *Статистическая зависимость между двумя переменными, при которой каждому значению одной переменной соответствует определенное условное математическое ожидание (среднее значение) другой, называется* ***корреляционной.*** Иначе, ***корреляционной зависимостью*** *между двумя переменными величинами называется функциональная зависимость между значениями одной из них и условным математическим ожиданием другой.*

Корреляционная зависимость может быть представлена в виде:

$$M_x(Y)=\varphi(x) \qquad (12.1) \quad \text{или} \quad M_y(X)=\psi(y). \qquad (12.2)$$

Предполагается, что $\varphi(x)\neq \text{const}$ и $\psi(y)\neq \text{const}$, т.е. если при изменении x или y условные математические ожидания $M_x(Y)$ и $M_y(X)$ не изменяются, то говорят, что корреляционная зависимость между переменными X и Y отсутствует.

Сравнивая различные виды зависимости между X и Y, можно сказать, что с изменением значений переменной X при функциональной зависимости однозначно изменяется определенное *значение* переменной Y, при корреляционной — определенное *среднее значение* (условное математическое ожидание) Y, а при статистической — определенное (условное) *распределение* переменной Y. Таким образом, из рассмотренных зависимостей *наиболее общей выступает статистическая зависимость*[2]. Каждая корреляционная зависимость является статистической, но не каждая статистическая зависимость является корреляционной. Функциональная зависимость представляет частный случай корреляционной (об этом речь еще пойдет ниже, в § 12.3).

Уравнения **(12.1)** и **(12.2)** называются *модельными уравнениями регрессии* (или просто *уравнениями регрессии*) соответственно Y по X и X по Y[1], функции $\varphi(x)$ и $\psi(y)$ — *модельными функциями регрессии* (или *функциями регрессии*), а их графики — *модельными линиями регрессии* (или *линиями регрессии*).

Для отыскания модельных уравнений регрессии, вообще говоря, необходимо знать *закон распределения двумерной случайной величины* (X,Y). На практике исследователь, как правило, располагает лишь *выборкой* пар значений (x_i,y_i) ограниченного объема. В этом случае речь может идти об оценке (приближенном выражении) по выборке функции регрессии. Такой наилучшей (в смысле метода наименьших квадратов) оценкой является *выборочная линия* (*кривая*) *регрессии* Y по X

$$y_x=\hat{\varphi}(x,b_0,b_1,...,b_p) \qquad (12.3)$$

где y_x — *условная* (*групповая*) *средняя* переменной Y при фиксированном значении переменной $X=x$; $b_0,b_1,...,b_p$ — параметры кривой.

Аналогично определяется *выборочная линия* (*кривая*) *регрессии* X по Y:

$$x_y=\hat{\psi}(y,c_0,c_1,...,c_p), \qquad (12.4)$$

где x_y — *условная* (*групповая*) *средняя* переменной X при фиксированном значении переменной $Y=y$; $c_0,c_1,...,c_p$ — параметры кривой.

Уравнения **(12.3)**, **(12.4)** называют также *выборочными уравнениями регрессии* соответственно Y по X и X по Y[2].

При правильно определенных аппроксимирующих функциях $\hat{\varphi}(x,b_0,b_1,...,b_p)$ и $\hat{\psi}(y,c_0,c_1,...,c_p)$ с увеличением объема выборки $(n\to\infty)$ они будут сходиться по вероятности соответственно к функциям регрессии $\varphi(x)$ и $\psi(y)$.

[1] Для условного математического ожидания в литературе используется также обозначение $M(Y|X=x)$.

[2] Хотя статистическая зависимость и является наиболее общей из рассмотренных, она не отражает любую возможную зависимость между переменными в условиях неопределенности. Например, можно предполагать, что существует некоторая зависимость между числом (продолжительностью) военных конфликтов и числом изобретений за определенный период времени. Эта зависимость, хотя и сводится к зависимости между событиями с неопределенным исходом (могут произойти или не произойти), но не является статистической, ибо каждому значению одной переменной нельзя поставить в соответствие распределение другой, так как к таким уникальным (единичным) и неповторяемым в одинаковых условиях событиям, какими являются соответственно военные конфликты и изобретения, неприменимо само понятие вероятности (см. § 1.3).

[1] Или Y на X и X на Y.

[2] В дальнейшем для краткости там, где это очевидно по смыслу, мы часто и выборочные уравнения (линии) регрессии будем называть просто уравнениями (линиями) регрессии.

Статистические связи между переменными можно изучать методами корреляционного и регрессионного анализа. *Основной задачей регрессионного анализа является установление формы и изучение зависимости между переменными. Основной задачей корреляционного анализа — выявление связи между случайными переменными и оценка ее тесноты.*

Вначале (§ 12.2, 12.3) познакомимся с основными понятиями корреляционного и регрессионного анализа, а затем (§ 12.4–12.7, 13.1–13.8) перейдем к более детальному изучению этих методов.

12.2. Линейная парная регрессия

Данные о статистической зависимости удобно задавать в виде *корреляционной таблицы.*

Рассмотрим в качестве примера зависимость между суточной выработкой продукции Y (т) и величиной основных производственных фондов X (млн руб.) для совокупности 50 однотипных предприятий (табл. 12.1).

Таблица 12.1

Величина ОПФ, млн. руб. (X)	*Середины интервалов*	*Суточная выработка продукции, т (Y)*					*Всего* n_i	*Групповая средняя, т* $\bar{y}_i$
		7–11	11–15	15–19	19–23	23–27		
	x_i \ y_j	9	13	17	21	25		
20–25	22,5	2	1	–	–	–	3	10,3
25–30	27,5	3	6	4	–	–	13	13,3
30–35	32,5	–	3	11	7	–	21	17,8
35–40	37,5	–	1	2	6	2	11	20,3
40–45	42,5	–	–	–	1	1	2	23,0
Всего n_j		5	11	17	14	3	50	–
Групповая средняя $\bar{x}_j$, млн руб.		25,5	29,3	31,9	35,4	39,2	–	–

(В таблице через x_i и y_j обозначены середины соответствующих интервалов, а n_i и n_j — соответственно их частоты).

Изобразим полученную зависимость графически точками координатной плоскости (рис. 12.1). Такое изображение статистической зависимости называется *полем корреляции.*

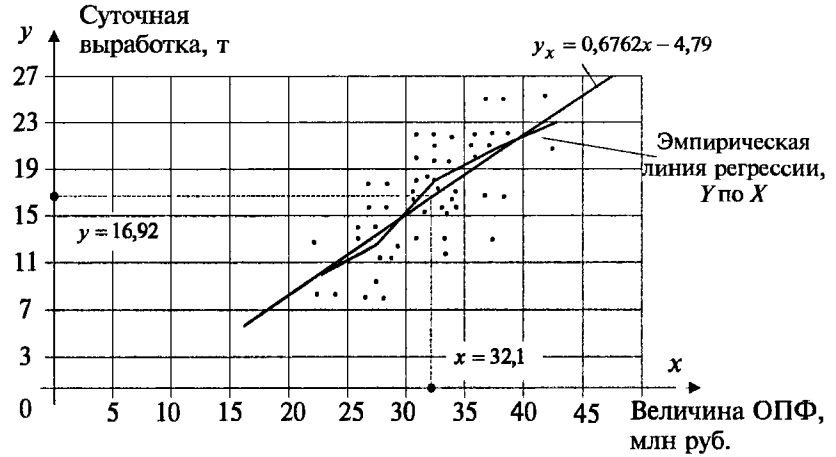


Рис. 12.1

Для каждого значения x_i ($i = 1,2,...,l$), т.е. для каждой строки корреляционной таблицы вычислим групповые средние

$$\bar{y}_i = \frac{\sum_{j=1}^{m} y_j n_{ij}}{n_i}, \qquad (12.5)$$

где n_{ij} — частоты пар (x_i, y_j) и $n_i = \sum_{j=1}^{m} n_{ij}$; m — число интервалов по переменной Y.

Вычисленные групповые средние $\bar{y}_i$ поместим в последнем столбце корреляционной таблицы и изобразим графически в виде ломаной, называемой *эмпирической линией регрессии* Y по X (рис. 12.1).

Аналогично для каждого значения y_j ($j = 1,2,...,m$) по формуле

$$\bar{x}_j = \frac{\sum_{i=1}^{l} x_i n_{ij}}{n_j} \qquad (12.6)$$

вычислим групповые средние $\bar{x}_j$ (см. нижнюю строку корреляционной таблицы)[1], где $n_j = \sum_{i=1}^{l} n_{ij}$, l — число интервалов по переменной X.

[1] Чтобы не загромождать чертеж, *эмпирическая линия регрессии* X по Y на рис. 12.1 не показана.

По виду ломаной можно предположить наличие линейной корреляционной зависимости Y по X между двумя рассматриваемыми переменными, которая графически выражается тем точнее, чем больше объем выборки (число рассматриваемых предприятий) n:

$$n = \sum_{i=1}^{l} n_i = \sum_{j=1}^{m} n_j = \sum_{i=1}^{l}\sum_{j=1}^{m} n_{ij}. \qquad (12.7)$$

Поэтому уравнение регрессии (12.3) будем искать в виде:

$$y_x = b_0 + b_1 x. \qquad (12.8)$$

Отвлечемся на время от рассматриваемого примера и найдем формулы расчета неизвестных параметров уравнения линейной регрессии.

С этой целью применим ***метод наименьших квадратов***, согласно которому *неизвестные параметры* b_0 *и* b_1 *выбираются таким образом, чтобы сумма квадратов отклонений эмпирических групповых средних* $\bar{y}_i$, *вычисленных по формуле* (12.5), *от значений* y_{x_i}, *найденных по уравнению регрессии* (12.8), *была минимальной*:

$$S = \sum_{i=1}^{l}\left(y_{x_i} - \bar{y}_i\right)^2 n_i = \sum_{i=1}^{l}\left(b_0 + b_1 x_i - \bar{y}_i\right)^2 n_i \to \min. \qquad (12.9)$$

На основании необходимого условия экстремума функции двух переменных $S = S(b_0, b_1)$ приравниваем к нулю ее частные производные, т.е.

$$\begin{cases} \dfrac{dS}{db_0} = 2\sum\limits_{i=1}^{l}\left(b_0 + b_1 x_i - \bar{y}_i\right) n_i = 0, \\ \dfrac{dS}{db_1} = 2\sum\limits_{i=1}^{l}\left(b_0 + b_1 x_i - \bar{y}_i\right) x_i n_i = 0, \end{cases}$$

откуда после преобразований получим *систему нормальных уравнений* для определения параметров линейной регрессии:

$$\begin{cases} b_0\sum\limits_{i=1}^{l} n_i + b_1\sum\limits_{i=1}^{l} x_i n_i = \sum\limits_{i=1}^{l} \bar{y}_i n_i, \\ b_0\sum\limits_{i=1}^{l} x_i n_i + b_1\sum\limits_{i=1}^{l} x_i^2 n_i = \sum\limits_{i=1}^{l} x_i \bar{y}_i n_i. \end{cases} \qquad (12.10)$$

Учитывая (12.5), преобразуем выражения:

$$\sum_{i=1}^{l} \bar{y}_i n_i = \sum_{i=1}^{l}\left(\frac{\sum\limits_{j=1}^{m} y_j n_{ij}}{n_i}\right) n_i = \sum_{i=1}^{l}\sum_{j=1}^{m} y_j n_{ij} = \sum_{j=1}^{m} y_j \sum_{i=1}^{l} n_{ij} = \sum_{j=1}^{m} y_j n_j,$$

$$\sum_{i=1}^{l} x_i \bar{y}_i n_i = \sum_{i=1}^{l} x_i \left(\frac{\sum\limits_{j=1}^{m} y_j n_{ij}}{n_i}\right) n_i = \sum_{i=1}^{l}\sum_{j=1}^{m} x_i y_j n_{ij}.$$

Теперь с учетом (12.7), разделив обе части уравнений (12.10) на n, получим систему нормальных уравнений в виде:

$$\begin{cases} b_0 + b_1\bar{x} = \bar{y}, \\ b_0\bar{x} + b_1\overline{x^2} = \overline{xy}, \end{cases} \qquad (12.11)$$

где соответствующие средние определяются по формулам:

$$\bar{x} = \frac{\sum\limits_{i=1}^{l} x_i n_i}{n}, \quad \bar{y} = \frac{\sum\limits_{j=1}^{m} y_j n_j}{n}, \qquad (12.12)$$

$$\overline{xy} = \frac{\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{m} x_i y_j n_{ij}}{n}, \qquad (12.13)$$

$$\overline{x^2} = \frac{\sum\limits_{i=1}^{l} x_i^2 n_i}{n}. \qquad (12.14)$$

Подставляя значение $b_0 = \bar{y} - b_1\bar{x}$ из первого уравнения системы (12.11) в уравнение регрессии (12.8), получим $y_x = \bar{y} - b_1\bar{x} + b_1 x$, или

$$y_x - \bar{y} = b_1\left(x - \bar{x}\right). \qquad (12.15)$$

Коэффициент b_1 в уравнении регрессии, называемый *выборочным коэффициентом регрессии* (или просто *коэффициентом регрессии*) Y по X, будем обозначать символом b_{yx}. Теперь уравнение регрессии Y по X запишется так:

$$y_x - \bar{y} = b_{yx}\left(x - \bar{x}\right). \qquad (12.16)$$

Коэффициент регрессии Y по X показывает, на сколько единиц в среднем изменяется переменная Y при увеличении переменной X на одну единицу.

Решая систему **(12.11)**, найдем

$$b_{yx} = b_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{s_x^2} = \frac{\mu}{s_x^2}, \quad \textbf{(12.17)}$$

где s_x^2 — выборочная дисперсия переменной X (см. **(8.10)**):

$$s_x^2 = \overline{x^2} - \bar{x}^2 = \frac{\sum_{i=1}^{l} x_i^2 n_i}{n} - (\bar{x})^2; \quad \textbf{(12.18)}$$

μ — выборочный корреляционный момент или выборочная ковариация[1]:

$$\mu = \overline{xy} - \bar{x}\,\bar{y} = \frac{\sum_{i=1}^{l}\sum_{j=1}^{m} x_i y_j n_{ij}}{n} - \bar{x}\,\bar{y}. \quad \textbf{(12.19)}$$

Рассуждая аналогично и полагая уравнение регрессии **(12.4)** линейным, можно привести его к виду:

$$x_y - \bar{x} = b_{xy}(y - \bar{y}), \quad \textbf{(12.20)}$$

где $$b_{xy} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{y^2} - \bar{y}^2} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{s_y^2} = \frac{\mu}{s_y^2} \quad \textbf{(12.21)}$$

— *выборочный коэффициент регрессии* (или просто *коэффициент регрессии*) *X по Y, показывающий, на сколько единиц в среднем изменяется переменная X при увеличении переменной Y на одну единицу*;

$$s_y^2 = \overline{y^2} - \bar{y}^2 = \frac{\sum_{j=1}^{m} y_j^2 n_j}{n} - (\bar{y})^2 \quad \textbf{(12.22)}$$

—выборочная дисперсия переменной Y.

[1] Для выборочной ковариации переменных X и Y используется также обозначение $\hat{\mathrm{cov}}(X, Y)$.

Так как числители в формулах **(12.17)** и **(12.21)** для b_{yx} и b_{xy} совпадают, а знаменатели — положительные величины, то *коэффициенты регрессии* b_{yx} и b_{xy} *имеют одинаковые знаки*, определяемые знаком μ. Из уравнений регрессии **(12.16)** и **(12.20)** следует, что коэффициенты b_{yx} и $1/b_{xy}$ определяют угловые коэффициенты (тангенсы углов наклона) к оси Ox соответствующих линий регрессии, пересекающихся в точке $(\bar{x}, \bar{y})$ (см. рис. 12.3).

▷ **Пример 12.1.** По данным табл. 12.1 найти уравнения регрессии Y по X и X по Y и пояснить их смысл.

Р е ш е н и е. Вычислим все необходимые суммы:

$$\sum_{i=1}^{l} x_i n_i = 22{,}5 \cdot 3 + 27{,}5 \cdot 13 + 32{,}5 \cdot 21 + 37{,}5 \cdot 11 + 42{,}5 \cdot 2 = 1605;$$

$$\sum_{i=1}^{l} x_i^2 n_i = 22{,}5^2 \cdot 3 + 27{,}5^2 \cdot 13 + 32{,}5^2 \cdot 21 + 37{,}5^2 \cdot 11 + 42{,}5^2 \cdot 2 = 52\,612{,}5;$$

$$\sum_{j=1}^{m} y_j n_j = 9 \cdot 5 + 13 \cdot 11 + 17 \cdot 17 + 21 \cdot 14 + 25 \cdot 3 = 846;$$

$$\sum_{j=1}^{l} y_j^2 n_j = 9^2 \cdot 5 + 13^2 \cdot 11 + 17^2 \cdot 17 + 21^2 \cdot 14 + 25^2 \cdot 3 = 15\,226;$$

$$\sum_{i=1}^{l}\sum_{j=1}^{m} x_i y_j n_{ij} = 22{,}5 \cdot 9 \cdot 2 + 22{,}5 \cdot 1 \cdot 13 + \ldots + 42{,}5 \cdot 1 \cdot 21 + 42{,}5 \cdot 1 \cdot 25 = 27\,895$$

(обходим все заполненные клетки корреляционной таблицы).

Затем по формулам **(12.12)**—**(12.22)** находим выборочные характеристики и параметры уравнений регрессии:

$\bar{x} = 1605/50 = 32{,}1$ (млн руб.); $\bar{y} = 846/50 = 16{,}92$ (т);

$s_x^2 = 52\,612{,}5/50 - 32{,}1^2 = 21{,}84$; $s_y^2 = 15\,226/50 - 16{,}92^2 = 18{,}2336$;

$\mu = 27\,895/50 - 32{,}1 \cdot 16{,}92 = 14{,}768$;

$b_{yx} = 14{,}768/21{,}84 = 0{,}6762$; $b_{xy} = 14{,}768/18{,}2336 = 0{,}8099$.

Итак, уравнения регрессии

$y_x - 16{,}92 = 0{,}6762(x - 32{,}1)$ или $y_x = 0{,}6762x - 4{,}79$,

$x_y - 32{,}1 = 0{,}8099(y - 16{,}92)$ или $x_y = 0{,}8099y + 18{,}40$.

Из первого уравнения регрессии Y по X (его график показан на рис. 12.1) следует, что при увеличении основных производственных фондов (ОПФ) X на 1 млн руб. суточная выработка продукции Y предприятия увеличивается в среднем на 0,6762 т. Второе уравнение регрессии X по Y показывает, что для увеличения суточной выработки продукции Y на 1 т необходимо в среднем увеличить ОПФ X на 0,8099 млн руб. (отметим, что свободные члены в уравнениях регрессии не имеют реального смысла). ▶

Параметры уравнений регрессии **(12.8)** могут быть вычислены **упрощенным способом** (аналогично тому, как вычислялись числовые характеристики вариационного ряда в § 8.4). С этой целью от значений переменных x_i и y_j переходят к новым значениям $u_i = \dfrac{x_i - c}{k}$ и $v_j = \dfrac{y_j - c'}{k'}$, где k и k' — величины интервалов, а c и c' — середины серединных интервалов соответственно по переменной X или Y. Тогда в соответствии с **(8.20)** и **(8.21)**

$$\overline{x} = \frac{\sum_{i=1}^{l} u_i n_i}{n} \cdot k + c, \qquad \textbf{(12.23)} \qquad \overline{y} = \frac{\sum_{j=1}^{m} v_j n_j}{n} \cdot k' + c', \qquad \textbf{(12.24)}$$

$$s_x^2 = \frac{\sum_{i=1}^{l} u_i^2 n_i}{n} \cdot k^2 - (\overline{x} - c)^2, \qquad \textbf{(12.25)}$$

$$s_y^2 = \frac{\sum_{j=1}^{m} v_j^2 n_j}{n} \cdot k'^2 - (\overline{y} - c')^2 \qquad \textbf{(12.26)}$$

Покажем, что в этом случае формула для ковариации μ **(12.19)** примет вид:

$$\mu = \frac{\sum_{i=1}^{l}\sum_{j=1}^{m} u_i v_j n_{ij}}{n} \cdot k \cdot k' - (\overline{x} - c)(\overline{y} - c'). \qquad \textbf{(12.27)}$$

□ Представим правую часть равенства **(12.27)** в виде:

$\overline{uv}kk' - (\overline{x} - c)(\overline{y} - c')$, где $\overline{uv} = \dfrac{\sum_{i=1}^{l}\sum_{j=1}^{m} u_i v_j n_{ij}}{n}$ — средняя арифметическая произведений вариантов

$$u_i v_j = \frac{x_i - c}{k} \cdot \frac{y_j - c'}{k'} = \frac{x_i y_j - c' x_i - c y_j + cc'}{kk'}.$$

Учитывая свойства средней,

$\overline{uv} = \dfrac{1}{kk'}\left(\overline{xy} - c'\overline{x} - c\overline{y} + cc'\right)$, откуда

$\overline{uv}kk' - (\overline{x} - c)(\overline{y} - c') = \overline{xy} - c'\overline{x} - c\overline{y} + cc' - (\overline{x} - c)(\overline{y} - c') =$

$= \overline{xy} - \overline{x}\,\overline{y} = \mu$ (по определению **(12.13)**). ■

▷ **Пример 12.2.** По данным табл. 12.1 найти упрощенным способом уравнения регрессии Y по X и X по Y и пояснить их смысл.

Р е ш е н и е. Возьмем постоянную k равной величине интервала по переменной X, т.е. $k = 5$, а постоянную c — равной середине серединного, третьего, интервала, т.е. $c = 32{,}5$. Аналогично по переменной Y $k' = 4$, $c' = 17$. Итак, $u_i = (x_i - 32{,}5)/5$; $v_j = (y_j - 17)/4$. Представим корреляционную табл. 12.1 в виде табл. 12.2.

Таблица 12.2

x_i \ y_j		9	13	17	21	25	n_i	$u_i n_i$	$u_i^2 n_i$	$\sum_{j=1}^{5} u_i v_j n_{ij}$
	v_j / u_i	−2	−1	0	1	2				
22,5	−2	2_4	1_2	—	—	—	3	−6	12	10
27,5	−1	3_2	6_1	4_0	—	—	13	−13	13	12
32,5	0	—	3_0	11_0	7_0	—	21	0	0	0
37,5	1	—	1_{-1}	2_0	6_1	2_2	11	11	11	9
42,5	2	—	—	—	1_2	1_4	2	4	8	6
n_j		5	11	17	14	3	50	−4	44	—
$v_j n_j$		−10	−11	0	14	6	−1	—	—	—
$v_j^2 n_j$		20	11	0	14	12	57	—	—	—
$\sum_{i=1}^{5} u_i v_j n_{ij}$		14	7	0	8	8	—	—	—	37

Вычислим необходимые суммы:

$$\sum_{i=1}^{5} u_i n_i = (-2)\cdot 3+(-1)\cdot 13+0\cdot 21+1\cdot 11+2\cdot 2=-4;$$

$$\sum_{i=1}^{5} u_i^2 n_i = (-2)^2\cdot 3+(-1)^2\cdot 13+0^2\cdot 21+1^2\cdot 11+2^2\cdot 2=44;$$

$$\sum_{j=1}^{5} v_j n_j = (-2)\cdot 5+(-1)\cdot 11+0\cdot 17+1\cdot 14+2\cdot 3=-1;$$

$$\sum_{j=1}^{5} v_j^2 n_j = (-2)^2\cdot 5+(-1)^2\cdot 11+0^2\cdot 17+1^2\cdot 14+2^2\cdot 3=57.$$

Для упрощения вычислений расчеты указанных сумм целесообразно проводить непосредственно в таблице (см. соответственно два предпоследних столбца и две предпоследние строки со значениями необходимых сумм в итоговых строке и столбце).

Для удобства вычисления суммы $\sum_{i=1}^{5}\sum_{j=1}^{5} u_i v_j n_{ij}$ вначале рассчитываем $u_i v_j$ и проставляем эти значения под соответствующими частотами, а затем находим произведения $(u_i v_j) n_{ij}$, которые суммируем по строке и столбцу, и записываем полученные числа соответственно в последнем столбце и последней строке табл. 12.2. Например, на пересечении первой строки и первого столбца табл. 12.2 получим 2_4, т.е. частота $n_{11} = 2$, $u_1 v_1 = (-2)(-2) = 4$, а $(u_1 v_1) n_{11} = 4\cdot 2 = 8$ и т.д. Итак, суммируя произведения $u_i v_j n_{ij}$ в последнем столбце или в последней строке, получим в правом нижнем углу табл. 12.2 $\sum_{i=1}^{5}\sum_{j=1}^{5} u_i v_j n_{ij} = 37$.

Теперь по формулам **(12.23)**—**(12.27)** имеем:

$$\bar{x} = \frac{-4}{50}\cdot 5+32{,}5 = 32{,}1 \text{ (млн руб.)};$$

$$\bar{y} = \frac{-1}{50}\cdot 4+17 = 16{,}92 \text{ (т)};$$

$$s_x^2 = \frac{44}{50}\cdot 5-(32{,}1-32{,}5)^2 = 21{,}84;$$

$$s_y^2 = \frac{57}{50}\cdot 4-(16{,}92-17)^2 = 18{,}2336;$$

$$\mu = \frac{37}{50}\cdot 5\cdot 4-(32{,}1-32{,}5)(16{,}92-17) = 14{,}768.$$

Далее уравнения регрессии находятся и интерпретируются так же, как в примере **12.1**. ▶

12.3. Коэффициент корреляции

Перейдем к оценке тесноты корреляционной зависимости. Рассмотрим наиболее важный для практики и теории случай *линейной зависимости* вида **(12.16)**.

На первый взгляд подходящим измерителем тесноты связи Y от X является коэффициент регрессии b_{yx} ибо, как уже отмечено, он показывает, на сколько единиц в среднем изменяется Y, когда X увеличивается на одну единицу. Однако b_{yx} зависит от единиц измерения переменных. Например, в полученной ранее зависимости он увеличится в 1000 раз, если величину основных производственных фондов X выразить не в млн руб., а в тыс. руб.

Очевидно, что для «исправления» b_{yx} как показателя тесноты связи нужна такая стандартная система единиц измерения, в которой данные по различным характеристикам оказались бы сравнимы между собой. Статистика знает такую систему единиц. Эта система использует в качестве единицы измерения переменной ее *среднее квадратическое отклонение* s.

Представим уравнение **(12.16)** в эквивалентном виде:

$$\frac{y_x-\bar{y}}{s_y} = \left(b_{yx}\frac{s_x}{s_y}\right)\frac{x-\bar{x}}{s_x}. \qquad \textbf{(12.28)}$$

В этой системе величина

$$r = b_{yx}\frac{s_x}{s_y} \qquad \textbf{(12.29)}$$

показывает, *на сколько величин s_y изменится в среднем Y, когда X увеличится на одно s_x.*

Величина r является показателем тесноты линейной связи и называется ***выборочным коэффициентом корреляции*** (или просто ***коэффициентом корреляции***).

На рис. 12.2 приведены две корреляционные зависимости переменной Y по X. Очевидно, что в случае *а*) зависимость меж-

ду переменными менее тесная и коэффициент корреляции должен быть меньше, чем в случае б), так как точки корреляционного поля а) дальше отстоят от линии регрессии, чем точки поля б).

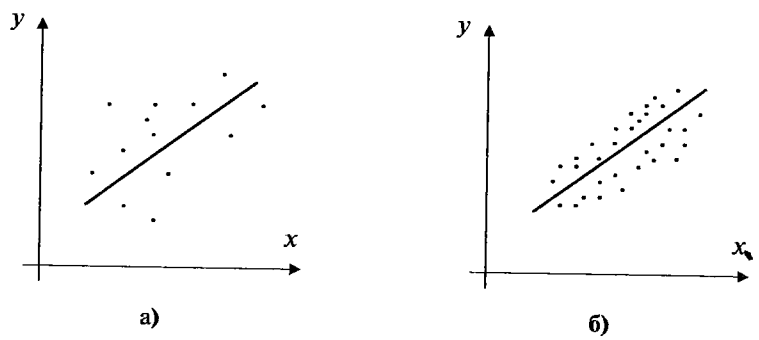


Рис. 12.2

Нетрудно видеть, что r совпадает по знаку с b_{yx} (а значит, и с b_{xy}). Если $r > 0$ ($b_{yx}>0$, $b_{xy}>0$), то корреляционная связь между переменными называется *прямой*, если $r < 0$ ($b_{yx}<0$, $b_{xy}<0$) — *обратной*. При прямой (обратной) связи увеличение одной из переменных ведет к увеличению (уменьшению) условной (групповой) средней другой.

Учитывая **(12.17)**, формулу для r представим в виде:

$$r=\frac{\overline{xy}-\bar{x}\,\bar{y}}{s_x s_y}. \qquad \textbf{(12.30)}$$

Отсюда видно, что формула для r симметрична относительно двух переменных, т.е. переменные X и Y можно менять местами. Тогда аналогично **(12.29)** можно записать:

$$r=b_{xy}\frac{s_y}{s_x}. \qquad \textbf{(12.31)}$$

Найдя произведение обеих частей равенств **(12.29)** и **(12.31)**, получим

$$r^2=b_{yx}b_{xy} \qquad \textbf{(12.32)}$$

или

$$r=\pm\sqrt{b_{yx}b_{xy}}, \qquad \textbf{(12.33)}$$

т.е. *коэффициент корреляции r переменных X и Y есть средняя геометрическая коэффициентов регрессии, имеющая их знак.*

▷ **Пример 12.3.** Вычислить коэффициент корреляции между величиной основных производственных фондов X и суточной выработкой продукции Y (по данным табл. 12.1).

Р е ш е н и е. Выше (см. примеры **12.1**, **12.2**) получили b_{yx}=0,6762 и b_{xy} = 0,8099. По формуле **(12.33)** $r=+\sqrt{0{,}6762\cdot 0{,}8099}=$ $=0{,}740$ (берем радикал со знаком +, так как коэффициенты b_{yx} и b_{xy} положительны). Итак, связь между рассматриваемыми переменными прямая и достаточно тесная (ибо r близок к 1)[1]. ▶

▷ **Пример 12.4.** При исследовании корреляционной зависимости между объемом валовой продукции Y (млн руб.) и среднесуточной численностью работающих X (тыс. чел.) для ряда предприятий отрасли получено следующее уравнение регрессии X по Y: x_y=0,2y − 2,5. Коэффициент корреляции между этими признаками оказался равным 0,8, а средний объем валовой продукции предприятий составил 40 млн руб. Найти: а) среднее значение среднесуточной численности работающих на предприятиях; б) уравнение регрессии Y по X; в) средний объем валовой продукции на предприятиях со среднесуточной численностью работающих 4 тыс. чел.

Р е ш е н и е. а) Обе линии регрессии Y по X и X по Y пересекаются в точке $(\bar{x},\bar{y})$, поэтому $\bar{x}$ найдем по заданному уравнению регрессии при $y=\bar{y}=40$, т.е. $\bar{x}=0{,}2\cdot 40-2{,}5=5{,}5$ (тыс. чел.).

б) Учитывая **(12.32)**, вычислим коэффициент регрессии b_{yx}: $b_{yx}=\dfrac{r^2}{b_{xy}}=\dfrac{0{,}8^2}{0{,}2}=3{,}2$. Теперь по формуле **(12.16)** получим уравнение регрессии Y по X: $y_x-40=3{,}2(x-5{,}5)$ или $y_x=3{,}2x+22{,}4$.

в) $y_{x=4}$ найдем по полученному уравнению регрессии Y по X: $y_{x=4}$=3,2 · 4+22,4=35,2 (млн руб.). ▶

Отметим другие модификации формулы r, полученные из **(12.30)** с помощью формул **(12.12)**—**(12.14)**, **(12.8)**, **(12.22)**:

$$r=\frac{\sum_{i=1}^{l}\sum_{j=1}^{m}(x_i-\bar{x})(y_j-\bar{y})n_{ij}}{ns_x s_y}; \qquad \textbf{(12.34)}$$

[1] См. ниже свойство **1** коэффициента корреляции.

$$r=\frac{n\sum_{i=1}^{l}\sum_{j=1}^{m}x_i y_j n_{ij}-\left(\sum_{i=1}^{l}x_i n_i\right)\left(\sum_{j=1}^{m}y_j n_j\right)}{\sqrt{n\sum_{i=1}^{l}x_i^2 n_i-\left(\sum_{i=1}^{l}x_i n_i\right)^2}\cdot\sqrt{n\sum_{j=1}^{m}y_j^2 n_j-\left(\sum_{j=1}^{m}y_j n_j\right)^2}} \qquad (12.35)$$

Для практических расчетов наиболее удобна формула (12.35), так как по ней r находится непосредственно из данных наблюдений и на величине r не скажутся округления данных, связанные с расчетом средних и отклонений от них.

Если данные не сгруппированы в виде корреляционной таблицы и представляют n пар чисел (x_i, y_i), то для вычисления коэффициентов регрессии и корреляции в соответствующих формулах следует взять $n_{ij}=n_i=n_j=1$, $j=i$, а $\sum_{i=1}^{l}\sum_{j=1}^{m}$ заменить на $\sum_{i=1}^{n}$.

▷ **Пример 12.5.** Найти коэффициент корреляции между производительностью труда Y (тыс. руб.) и энерговооруженностью труда X (кВт) (в расчете на одного работающего) для 14 предприятий региона по следующим данным:

Таблица 12.3

x_i	2,8	2,2	3,0	3,5	3,2	3,7	4,0	4,8	6,0	5,4	5,2	5,4	6,0	9,0
y_i	6,7	6,9	7,2	7,3	8,4	8,8	9,1	9,8	10,6	10,7	11,1	11,8	12,1	12,4

Р е ш е н и е. Вычислим необходимые суммы:

$$\sum_{i=1}^{14}x_i=2{,}8+2{,}2+\ldots+6{,}0+9{,}0=64{,}2;$$

$$\sum_{i=1}^{14}x_i^2=2{,}8^2+2{,}2^2+\ldots+6{,}0^2+9{,}0^2=335{,}26;$$

$$\sum_{i=1}^{14}y_i=6{,}7+6{,}9+\ldots+12{,}1+12{,}4=132{,}9;$$

$$\sum_{i=1}^{14}y_i^2=6{,}7^2+6{,}9^2+\ldots+12{,}1^2+12{,}4^2=1313{,}95;$$

$$\sum_{i=1}^{14}x_i y_i=2{,}8\cdot6{,}7+2{,}2\cdot6{,}9+\ldots+6{,}0\cdot12{,}1+9{,}0\cdot12{,}4=650{,}99.$$

По формуле (12.35), полагая $n_{ij}=n_i=n_j=1$, $j=i$ и заменяя $\sum_{i=1}^{l}\sum_{j=1}^{m}$ на $\sum_{i=1}^{n}$, получим

$$r=\frac{n\sum_{i=1}^{n}x_i y_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum x_i\right)^2}\cdot\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}}= \qquad (12.35')$$

$$=\frac{14\cdot650{,}99-64{,}2\cdot132{,}9}{\sqrt{14\cdot335{,}26-64{,}2^2}\sqrt{14\cdot1313{,}95-132{,}4^2}}=0{,}898,$$

что говорит о тесной связи между переменными[1]. ▶

Отметим основные **свойства коэффициента корреляции** (при достаточно большом объеме выборки n), аналогичные свойствам коэффициента корреляции двух случайных величин (§ 5.6).

1. *Коэффициент корреляции принимает значения на отрезке* $[-1,1]$, т.е.

$$-1\le r\le 1. \qquad (12.36)$$

В зависимости от того, насколько $|r|$ приближается к 1, различают связь слабую, умеренную, заметную, достаточно тесную, тесную и весьма тесную, т.е. чем ближе $|r|$ к 1, тем теснее связь.

2. *Если все значения переменных увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то величина коэффициента корреляции не изменится.*

3. *При* $r=\pm1$ *корреляционная связь представляет* ***линейную функциональную зависимость****. При этом линии регрессии Y по X и X по Y совпадают и все наблюдаемые значения располагаются на общей прямой.*

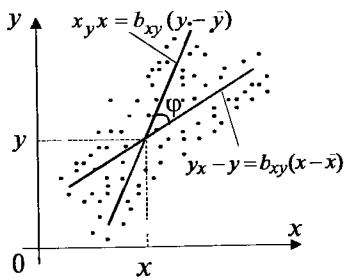


Рис. 12.3

□ Найдем $\operatorname{tg}\varphi$ между двумя прямыми регрессии (рис. 12.3) с угловыми коэффициентами $k_1=b_{yx}$ и

[1] См. ниже свойство **1** коэффициента корреляции.

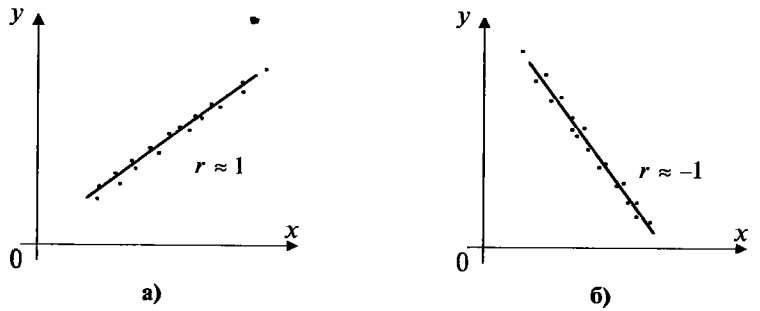
$k_2 = \frac{1}{b_{xy}}$, используя соответствующую формулу аналитической геометрии:

$$\operatorname{tg}\varphi = \frac{k_2 - k_1}{1 + k_2 k_1} = \frac{1 - b_{yx} b_{xy}}{b_{xy} + b_{yx}},$$

откуда с учетом **(12.24)** и **(12.26)**

$$\operatorname{tg}\varphi = \frac{1 - r^2}{r} \cdot \frac{s_x s_y}{s_x^2 + s_y^2}. \blacksquare \qquad \textbf{(12.37)}$$

Из полученной формулы видно, что чем теснее связь и чем ближе $|r|$ к 1, тем меньше угол φ между прямыми регрессии (уже образуемые ими «ножницы»), а при $r = \pm 1$ $\operatorname{tg}\varphi = \varphi = 0$ и линии регрессии сливаются (рис. 12.4 *а* и *б*).



а) б)

Рис. 12.4

4. *При $r = 0$ линейная корреляционная связь отсутствует. При этом групповые средние переменных совпадают с их общими средними, а линии регрессии Y по X и X по Y параллельны осям координат.*

□ Если $r = 0$, то коэффициент $b_{yx} = b_{xy} = 0$, и линии регрессии **(12.16)** и **(12.20)** имеют вид: $y_x = \bar{y}$ и $x_y = \bar{x}$ (рис. 12.5). ■

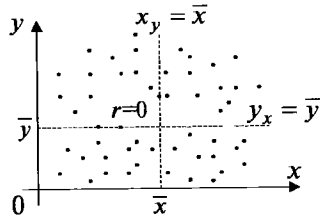
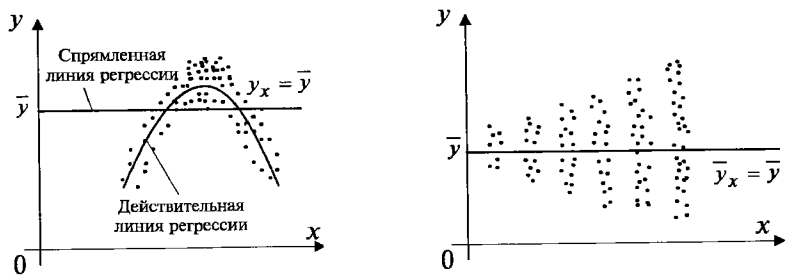


Рис. 12.5

Равенство $r = 0$ говорит лишь об отсутствии линейной корреляционной зависимости (некоррелированности переменных), но не вообще об отсутствии корреляционной, а тем более статистической, зависимости.

Так, например, для зависимостей, представленных на рис. 12.6 *а* и *б*, $r = 0$ и линии регрессии *Y* по *X* параллельны оси абсцисс. Однако по расположению точек корреляционного поля отчетливо просматривается взаимосвязь между переменными, отличная от линейной корреляционной. Так, в случае *а* — это нелинейная корреляционная (почти функциональная) зависимость; в случае *б* — статистическая зависимость, проявляющаяся в данном случае в том, что с изменением *x* групповые средние y_x не меняются, а меняется лишь рассеяние точек поля относительно линии регрессии.

Выборочный коэффициент корреляции *r* является *оценкой генерального коэффициента корреляции* ρ (о котором речь пойдет дальше), тем более точной, чем больше объем выборки *n*. И указанные выше свойства, строго говоря, справедливы для ρ. Однако при достаточно большом *n* их можно распространить и на *r*.



а) б)

Рис. 12.6

12.4. Основные положения корреляционного анализа. Двумерная модель

Корреляционный анализ (корреляционная модель) *— метод, применяемый тогда, когда данные наблюдений или эксперимента*

можно считать случайными и выбранными из совокупности, распределенной по многомерному нормальному закону.

Основная задача корреляционного анализа, как отмечено выше, *состоит в выявлении связи между случайными переменными путем точечной и интервальной оценок различных (парных, множественных, частных) коэффициентов корреляции.* Дополнительная задача корреляционного анализа (являющаяся основной в регрессионном анализе) заключается в оценке уравнений регрессии одной переменной по другой.

Рассмотрим простейшую модель корреляционного анализа — двумерную. Плотность совместного нормального распределения двух переменных X и Y имеет вид (см. § 5.7):

$$\varphi_N(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-L(x,y)}, \qquad \textbf{(12.38)}$$

где $L(x,y)=-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-a_x}{\sigma_x}\right)^2-2\rho\frac{x-a_x}{\sigma_x}\cdot\frac{y-a_y}{\sigma_y}+\left(\frac{y-a_y}{\sigma_y}\right)^2\right]$;

a_x, a_y — математические ожидания переменных X и Y;

σ_x^2, σ_y^2 — дисперсии переменных X и Y;

ρ — коэффициент корреляции между переменными X и Y, определяемый через корреляционный момент (ковариацию) K_{xy} по формуле **(5.38)**:

$$\rho=\frac{K_{xy}}{\sigma_x\sigma_y}=\frac{M[(X-a_x)(Y-a_y)]}{\sigma_x\sigma_y}, \qquad \textbf{(12.39)}$$

или с учетом **(5.40)**

$$\rho=\frac{M(XY)-a_xa_y}{\sigma_x\sigma_y}. \qquad \textbf{(12.40)}$$

Величина ρ характеризует тесноту связи между случайными переменными X и Y. Указанные пять параметров a_x, a_y, σ_x^2, σ_y^2, ρ дают исчерпывающие сведения о корреляционной зависимости между переменными.

В § 5.7 показано, что при совместном нормальном законе распределения случайных величин X и Y **(12.38)** выражения для

условных математических ожиданий, т.е. модельные уравнения регрессии **(12.1)** и **(12.2)**, выражаются линейными функциями:

$$M_x(Y)=a_y+\rho\frac{\sigma_y}{\sigma_x}(x-a_x), \qquad \textbf{(12.41)}$$

$$M_y(X)=a_x+\rho\frac{\sigma_x}{\sigma_y}(y-a_y). \qquad \textbf{(12.42)}$$

Из свойств коэффициента корреляции (§ 5.6) следует, что ρ *является показателем тесноты связи лишь в случае линейной зависимости (линейной регрессии) между двумя переменными, получаемой, в частности, в соответствии с* **(12.41)**, **(12.42)** *при их совместном нормальном распределении.*

Из § 5.6 также следует (см. формулы **(5.50)**, **(5.52)**), что условные дисперсии равны:

$$\sigma_y^2(Y)=\sigma_y^2(1-\rho^2), \quad \sigma_y^2(X)=\sigma_x^2(1-\rho^2),$$

т.е. степень рассеяния значений Y (или X) относительно линии регрессии Y по X (или X по Y) определяется двумя факторами: дисперсией σ_y^2 (σ_x^2) переменной Y (X) и коэффициентом корреляции ρ и не зависит от значений независимой переменной x (y). По мере приближения $|\rho|$ к 1 условная дисперсия $\sigma_x^2(Y)$ $(\sigma_y^2(X))\to 0$, и значения переменных все менее рассеяны относительно соответствующих линий регрессии, т.е. очевиден смысл коэффициента корреляции как показателя тесноты линейной корреляционной зависимости.

Генеральная совокупность в определенном смысле аналогична понятию случайной величины и ее закону распределения (см. § 9.1), поэтому для вышеназванных параметров используется и другая терминология: a_x, a_y (или $\bar{x}_0$, $\bar{y}_0$) — *генеральные средние*; σ_x^2, σ_y^2 — *генеральные дисперсии*; K_{xy} и ρ — *генеральные вариация* и *коэффициент корреляции.*

Для *оценки* генерального коэффициента корреляции ρ и модельных уравнений регрессии по выборке в формулах **(12.40)**—**(12.42)** необходимо заменить параметры a_x, a_y, σ_x^2, σ_y^2, K_{xy} их состоятельными выборочными оценками — соответственно

$\bar{x}$, $\bar{y}$ **(12.12)**, s_x^2 **(12.18)**, s_y^2 **(12.22)**, μ **(12.19)**. В этом случае получим знакомые нам формулы для определения выборочного коэффициента корреляции r **(12.30)** и выборочных уравнений регрессии **(12.16)**, **(12.20)**. Выше (§ 12.2 и 12.3) те же формулы получены иначе — на основе применения метода наименьших квадратов. Совпадение результатов объясняется некоторыми ценными свойствами оценок метода наименьших квадратов.

В § 12.3 мы ввели выборочный коэффициент корреляции r и рассмотрели его свойства, исходя из оценки близости точек корреляционного поля к прямой регрессии без учета предпосылок корреляционного анализа. Однако если эти предпосылки нарушаются (совместный закон распределения переменных не является нормальным, одна из исследуемых переменных не является случайной и т.п.), то r не следует рассматривать как строгую меру взаимосвязи переменных.

12.5. Проверка значимости и интервальная оценка параметров связи

В практических исследованиях о тесноте корреляционной зависимости между рассматриваемыми переменными судят фактически не по величине генерального коэффициента корреляции ρ (который обычно неизвестен), а по величине его выборочного аналога r. Так как r вычисляется по значениям переменных, случайно попавшим в выборку из генеральной совокупности, то в отличие от параметра ρ оценка r — *величина случайная.*

Пусть вычисленное значение $r \neq 0$. Возникает вопрос, объясняется ли это действительно существующей линейной корреляционной связью между переменными X и Y в генеральной совокупности или является следствием случайности отбора переменных в выборку (т.е. при другом отборе возможно, например, $r = 0$ или изменение знака r).

Обычно в этих случаях проверяется гипотеза H_0 об отсутствии линейной корреляционной связи между переменными в генеральной совокупности, т.е. H_0: $\rho = 0$. При справедливости этой гипотезы статистика

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \textbf{(12.43)}$$

имеет t-распределение Стьюдента с $k = n-2$ степенями свободы. Поэтому гипотеза H_0 отвергается, т.е. выборочный коэффициент корреляции r значимо (существенно) отличается от нуля, если

$$|t| = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} > t_{1-\alpha;\,k}, \qquad \textbf{(12.44)}$$

где $t_{1-\alpha;k}$ — табличное значение t-критерия Стьюдента, определенное на уровне значимости α при числе степеней свободы $k = n-2$.

▷ **Пример 12.6.** Проверить на уровне $\alpha = 0{,}05$ значимость коэффициента корреляции между переменными X и Y по данным табл. 12.1.

Р е ш е н и е. В примере **12.3** вычислен $r = 0{,}740$. Статистика критерия по **(12.43)**:

$$t = \frac{0{,}740\sqrt{50-2}}{\sqrt{1-0{,}740^2}} = 7{,}62.$$

Для уровня значимости $\alpha = 0{,}05$ и числа степеней свободы $k = 50 - 2 = 48$ находим критическое значение статистики $t_{0{,}95;48} = 2{,}01$ (см. табл. IV приложений). Поскольку $t > t_{0{,}95;48}$, коэффициент корреляции между суточной выработкой продукции Y и величиной основных производственных фондов X значимо отличается от нуля. ◀

Для значимого коэффициента корреляции r целесообразно найти *доверительный интервал* (*интервальную оценку*), который с заданной надежностью $\gamma = 1 - \alpha$ содержит (точнее, «накрывает») неизвестный генеральный коэффициент корреляции ρ. Для построения такого интервала необходимо знать выборочное распределение коэффициента корреляции r, которое при $\rho \neq 0$ несимметрично и очень медленно (с ростом n) сходится к нормальному распределению. Поэтому прибегают к специально подобранным функциям от r, которые сходятся к хорошо изученным распределениям. Чаще всего для подбора функции применяют *z-преобразование Фишера*:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}. \qquad \textbf{(12.45)}$$

Распределение z уже при небольших n является приближенно нормальным с математическим ожиданием

$$M(z)=\frac{1}{2}\ln\frac{1+\rho}{1-\rho}+\frac{\rho}{2(n-1)} \qquad \textbf{(12.46)}$$

и дисперсией

$$\sigma_z^2=\frac{1}{n-3}. \qquad \textbf{(12.47)}$$

Поэтому вначале строят доверительный интервал для $M(z)$:

$$z-t_{1-\alpha}\frac{1}{\sqrt{n-3}}\le M(z)\le z+t_{1-\alpha}\frac{1}{\sqrt{n-3}}, \qquad \textbf{(12.48)}$$

где $t_{1-\alpha}$ — нормированное отклонение z, определяемое с помощью функции Лапласа:

$$\Phi(t_{1-\alpha})=\gamma=1-\alpha. \qquad \textbf{(12.49)}$$

При определении границ доверительного интервала для ρ, т.е. для перехода от z к ρ, существует специальная таблица. При ее отсутствии переход может быть осуществлен по формуле:

$$r=\operatorname{th}z=\frac{e^z-e^{-z}}{e^z+e^{-z}}, \qquad \textbf{(12.50)}$$

где $\operatorname{th}z$ — гиперболический тангенс z.

Если коэффициент корреляции значим, то коэффициенты регрессии b_{yx} и b_{xy} также значимо отличаются от нуля, а интервальные оценки для соответствующих *генеральных коэффициентов регрессии* β_{yx} и β_{xy} могут быть получены по формулам, основанным на том, что статистики $(b_{yx}-\beta_{yx})/s_{b_{yx}}$, $(b_{xy}-\beta_{xy})/s_{b_{xy}}$ имеют t-распределение Стьюдента с $(n-2)$ степенями свободы:

$$b_{yx}-t_{1-\alpha;n-2}\frac{s_y\sqrt{1-r^2}}{s_x\sqrt{n-2}}\le\beta_{yx}\le b_{yx}+t_{1-\alpha;n-2}\cdot\frac{s_y\sqrt{1-r^2}}{s_x\sqrt{n-2}}; \qquad \textbf{(12.51)}$$

$$b_{xy}-t_{1-\alpha;n-2}\frac{s_x\sqrt{1-r^2}}{s_y\sqrt{n-2}}\le\beta_{xy}\le b_{xy}+t_{1-\alpha;n-2}\cdot\frac{s_x\sqrt{1-r^2}}{s_y\sqrt{n-2}}. \qquad \textbf{(12.51′)}$$

Z-преобразование Фишера может быть применено при проверке различных гипотез относительно коэффициента корреляции.

Например, если по данным выборки объема n вычислен коэффициент корреляции r, то для проверки нулевой гипотезы H_0 о том, что генеральный коэффициент корреляции ρ равен значению ρ_0, т.е. H_0: $\rho=\rho_0$, используется статистика

$$t=\frac{z(r)-z(\rho_0)}{\sqrt{\frac{1}{n-3}}}. \qquad \textbf{(12.52)}$$

А для проверки существенности (значимости) различия двух коэффициентов корреляции r_1 и r_2, полученных по выборкам объемов n_1 и n_2, т.е. для проверки гипотезы H_0: $\rho_1=\rho_2$, применяется статистика

$$t=\frac{z(r_1)-z(r_2)}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}}. \qquad \textbf{(12.52′)}$$

При достаточных объемах выборки (бóльших 10) можно считать, что при выполнении соответствующих нулевых гипотез статистики **(12.52)** и **(12.52′)** имеют приближенно нормальный закон распределения. Поэтому (см. § 10.6) гипотеза H_0 отвергается на уровне значимости α, если $|t|>t_{1-\alpha}$ (при использовании двустороннего критерия) или $|t|>t_{1-2\alpha}$ при использовании одностороннего критерия).

▷ **Пример 12.7.** По данным табл. 12.1 найти с надежностью 0,95 интервальные оценки (доверительные интервалы) параметров связи между суточной выработкой продукции Y и величиной основных производственных фондов X.

Р е ш е н и е. Так как коэффициент корреляции X и Y значим (см. пример **12.5**), то построим доверительный интервал для генерального коэффициента корреляции ρ, применяя z-преобразование Фишера. По **(12.45)**

$$z=\frac{1}{2}\ln\frac{1+0{,}740}{1-0{,}740}=0{,}9505.$$

По **(12.49)** из условия $\Phi(t_{1-\alpha})=0{,}95$ по таблице функции Лапласа находим $t_{0,95}$=1,96. По **(12.48)** построим доверительный интервал для $M(z)$:

$$0{,}9505-1{,}96\frac{1}{\sqrt{50-3}}\le M(z)\le 0{,}9505+1{,}96\frac{1}{\sqrt{50-3}}$$

или $0{,}6646\le M(z)\le 1{,}2364$. Находим границы доверительного интервала для ρ, используя специальную таблицу или формулу **(12.50)**: th 0,6646 <ρ< th 1,2364 или $0{,}581\le\rho\le 0{,}844$. В указанных границах на уровне значимости 0,05 (с надежностью 0,95) заключен генеральный коэффициент корреляции ρ.

Теперь построим доверительные интервалы для генеральных коэффициентов регрессии β_{yx} и β_{xy}. Вначале определим средние квадратические отклонения переменных:

$$s_x=\sqrt{s_x^2}=\sqrt{21{,}84}=4{,}673;\qquad s_y=\sqrt{s_y^2}=\sqrt{18{,}2336}=4{,}270;$$

Теперь по **(12.51)**:

$$0{,}6762-2{,}01\cdot\frac{4{,}270\cdot\sqrt{1-0{,}740^2}}{4{,}673\cdot\sqrt{50-2}}\le\beta_{yx}\le 0{,}6762+2{,}01\cdot\frac{4{,}270\cdot\sqrt{1-0{,}740^2}}{4{,}673\cdot\sqrt{50-2}}$$

или $0{,}4979\le\beta_{yx}\le 0{,}8545$. Аналогично по **(12.51′)**:

$0{,}5963\le\beta_{xy}\le 1{,}0235$. ▶

При содержательной интерпретации параметров ρ, β_{yx} и β_{xy} следует считаться в первую очередь с их *интервальными* (а не только точечными) оценками.

▷ **Пример 12.7а.** При исследовании связи между производительностью труда и уровнем механизации работ на предприятиях одной отрасли промышленности, расположенных в двух различных районах страны, вычислены коэффициенты корреляции $r_1=0{,}95$ и $r_2=0{,}88$ по выборкам объемов соответственно $n_1=14$ и $n_2=20$. Выяснить, имеются ли на уровне $\alpha=0{,}05$ существенные различия в тесноте связи между рассматриваемыми переменными на предприятиях отрасли в этих районах.

Р е ш е н и е. Проверяемая гипотеза H_0: $\rho_1=\rho_2$. В качестве альтернативной возьмем гипотезу H_0: $\rho_1\ne\rho_2$, т.е. применяем двусторонний критерий. По формуле **(12.51′)** с учетом **(12.45)** статистика

$$t=\frac{z(0{,}95)-z(0{,}88)}{\sqrt{\frac{1}{14-3}+\frac{1}{20-3}}}=\frac{1{,}832-1{,}376}{\sqrt{0{,}150}}=1{,}18.$$

Так как $t<t_{0{,}95}=1{,}96$, то гипотеза H_0 не отвергается, т.е. нет оснований считать существенным различие показателей связи между рассматриваемыми переменными на предприятиях двух районов страны. ▶

12.6. Корреляционное отношение и индекс корреляции

Введенный выше коэффициент корреляции, как уже отмечено, является полноценным показателем тесноты связи лишь в случае линейной зависимости между переменными. Однако часто возникает необходимость в достоверном показателе интенсивности связи при л ю б о й ф о р м е зависимости

Для получения такого показателя вспомним правило сложения дисперсий **(8.12)**:

$$s_y^2=s_{iy}'^2+\delta_{iy}^2,\qquad\textbf{(12.53)}$$

где s_y^2 — общая дисперсия переменной

$$s_y^2=\frac{\sum_{j=1}^{m}(y_j-\bar y)^2 n_i}{n},\qquad\textbf{(12.54)}$$

$s_{iy}'^2$ — средняя групповых дисперсий s_{iy}^2, или остаточная дисперсия —

$$s_{iy}'^2=\frac{\sum_{i=1}^{l}s_{iy}^2 n_i}{n},\qquad\textbf{(12.55)}$$

$$s_{iy}^2=\frac{\sum_{j=1}^{m}(y_j-\bar y_i)^2}{n},\qquad\textbf{(12.56)}$$

δ_{iy}^2 — межгрупповая дисперсия

$$\delta_{iy}^2=\frac{\sum_{i=1}^{l}(\bar y_i-\bar y)^2 n_i}{n}.\qquad\textbf{(12.57)}$$

Остаточной дисперсией измеряют ту часть колеблемости Y, которая возникает из-за изменчивости неучтенных факторов,

не зависящих от X. Межгрупповая дисперсия выражает ту часть вариации Y, которая обусловлена изменчивостью X. Величина

$$\eta_{yx} = \sqrt{\frac{\delta_{iy}^2}{s_y^2}} \qquad (12.58)$$

получила название *эмпирического корреляционного отношения* Y по X. Чем теснее связь, тем большее влияние на вариацию переменной Y оказывает изменчивость X по сравнению с неучтенными факторами, тем выше η_{yx}. Величина η_{yx}^2, называемая *эмпирическим коэффициентом детерминации*, показывает, какая часть общей вариации Y обусловлена вариацией X. Аналогично вводится *эмпирическое корреляционное отношение X по Y*:

$$\eta_{yx} = \sqrt{\frac{\delta_{ix}^2}{s_x^2}}. \qquad (12.59)$$

Отметим **основные свойства корреляционных отношений**[1] (при достаточно большом объеме выборки n):

1. *Корреляционное отношение есть неотрицательная величина, не превосходящая* 1: $0 \le \eta \le 1$.

2. *Если* $\eta = 0$, *то корреляционная связь отсутствует.*

3. *Если* $\eta = 1$, *то между переменными существует функциональная зависимость.*

4. $\eta_{yx} \ne \eta_{xy}$, т.е. в отличие от коэффициента корреляции r (для которого $r_{yx} = r_{xy} = r$) при вычислении корреляционного отношения существенно, какую переменную считать независимой, а какую — зависимой.

Эмпирическое корреляционное отношение η_{yx} *является показателем рассеяния точек корреляционного поля относительно эмпирической линии регрессии*, выражаемой ломаной, соединяющей значения y_i. Однако в связи с тем, что закономерное изменение y_i нарушается случайными зигзагами ломаной, возникающими вследствие остаточного действия неучтенных факторов, η_{yx} преувеличивает тесноту связи. Поэтому наряду с η_{yx} рассматривается показатель тесноты связи R_{yx}, характеризующий *рассеяние точек корреляционного поля относительно линии регрессии y_x*

[1] Эти свойства справедливы как для эмпирических корреляционных отношений η, так и для теоретических — R (см. ниже).

(**12.3**). Показатель R_{yx} получил название ***теоретического корреляционного отношения*** или ***индекса корреляции*** *Y по X*:

$$R_{yx} = \sqrt{\frac{\delta_y^2}{s_y^2}} = \sqrt{1 - \frac{s_y'^2}{s_y^2}}, \qquad (12.60)$$

где дисперсии δ_y^2 и $s_y'^2$ определяются по формулам (**12.54**)—(**12.56**), в которых групповые средние y_i заменены условными средними y_{x_i}, вычисленными пс уравнению регрессии (**12.16**).

Подобно R_{yx} вводится и *индекс корреляции X по Y*:

$$R_{xy} = \sqrt{\frac{\delta_x^2}{s_x^2}} = \sqrt{1 - \frac{s_x'^2}{s_x^2}}. \qquad (12.61)$$

Достоинством рассмотренных показателей η и R является то, что они могут быть вычислены при любой форме связи между переменными. Хотя η и завышает тесноту связи по сравнению с R, но для его вычисления не нужно знать уравнение регрессии. Корреляционные отношения η и R связаны с коэффициентом корреляции r следующим образом:

$$0 \le |r| \le R \le \eta \le 1. \qquad (12.62)$$

Покажем, что в случае л и н е й н о й модели (**12.3**), т.е. зависимости $\bar{y}_x - \bar{y} = b_{yx}(x - \bar{x})$, *индекс корреляции R_{yx} равен коэффициенту корреляции r (по абсолютной величине)*: $R_{yx} = |r|$ (или $R_{xy} = |r|$).

□ Полагаем для простоты $n_i = 1$ ($i = 1,2,..., l$).

По формуле (**12.60**)

$$R_{yx} = \sqrt{\frac{\delta_y^2}{s_y^2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_{x_i} - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \sqrt{\frac{\sum_{i=1}^{n} b_{yx}^2 (x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = |b_{yx}| \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 / n}{\sum_{i=1}^{n}(y_i - \bar{y})^2 / n}}$$

(так как из уравнения регрессии $y_{x_i} - \bar{y} = b_{yx}(x_i - \bar{x})$).

Теперь, учитывая формулы дисперсии, коэффициентов регрессии (**12.17**) и корреляции (**12.30**), получим:

$$R_{yx}=\frac{\left|\overline{xy}-\bar{x}\,\bar{y}\right|}{s_x^2}\sqrt{\frac{s_x^2}{s_y^2}}=\frac{\left|\overline{xy}-\bar{x}\,\bar{y}\right|}{s_x s_y}=|r|. \blacksquare$$

Коэффициент детерминации R^2, равный квадрату индекса корреляции (для парной линейной модели — r^2), *показывает долю общей вариации зависимой переменной, обусловленной регрессией или изменчивостью объясняющей переменной.*

Чем ближе R^2 к единице, тем лучше регрессия аппроксимирует эмпирические данные, тем теснее наблюдения примыкают к линии регрессии. Если $R^2 = 1$, то эмпирические точки (x, y) лежат на линии регрессии (см. рис. 12.4) и между переменными Y и X существует линейная функциональная зависимость. Если $R^2 = 0$, то вариация зависимой переменной полностью обусловлена воздействием неучтенных в модели переменных, и линия регрессии параллельна оси абсцисс (рис. 12.5).

Расхождение между η^2 и R^2 (или r^2) может быть использовано для проверки линейности корреляционной зависимости (см. ниже пример **12.10**).

Проверка значимости корреляционного отношения η основана на том, что статистика

$$F=\frac{\eta^2(n-m)}{(1-\eta^2)(m-1)} \qquad \textbf{(12.63)}$$

(где m — число интервалов по группировочному признаку) имеет F-распределение Фишера—Снедекора с $k_1 = m - 1$ и $k_2 = n - m$ степенями свободы. Поэтому η значимо отличается от нуля, если $F > F_{\alpha,k_1,k_2}$, где F_{α,k_1,k_2} — табличное значение F-критерия на уровне значимости α при числе степеней свободы $k_1 = m - 1$ и $k_2 = n - m$.

Индекс корреляции R двух переменных значим, если значение статистики

$$F=\frac{R^2(n-2)}{1-R^2} \qquad \textbf{(12.64)}$$

больше табличного F_{α,k_1,k_2}, где $k_1 = 1$ и $k_2 = n - 2$.

▷ **Пример 12.8.** По данным табл. 12.1 вычислить корреляционное отношение η_{yx} и индекс корреляции R_{yx} и проверить их значимость на уровне $\alpha = 0{,}05$.

Р е ш е н и е. Вначале определим η_{yx}. Ранее вычислены: общая средняя $\bar{y} = 16{,}92$, дисперсия $s_y^2 \approx 18{,}23$ (пример **12.2**), групповые средние $\bar{y}_i$ (табл. 12.1).

Частоты интервалов n_i указаны в предпоследней графе той же таблицы. Для удобства расчеты представим в табл. 12.4.

Таблица 12.4

x_i	n_i	$\bar{y}_i$	$(\bar{y}_i-\bar{y})^2 n_i$	y_{x_i}	$(\bar{y}_{x_i}-\bar{y})^2 n_i$
22,5	3	10,3	131,5	10,4	127,5
27,5	13	13,3	170,4	13,8	126,5
32,5	21	17,8	16,3	17,2	1,6
37,5	11	20,3	125,7	20,6	149,0
42,5	2	23,0	73,9	23,9	97,4
	Σ		517,8	—	502,0

Теперь по (**12.57**) $\delta_{iy}^2 = 517{,}8/50 = 10{,}36$ и по (**12.58**) $\eta_y=\sqrt{\dfrac{10{,}36}{18{,}23}}=\sqrt{0{,}568}=0{,}754$. Значение η_{yx} близко к величине $r = 0{,}740$ (полученной ранее в примере **12.3**). Поэтому оправдано сделанное выше на основании графического изображения эмпирической линии (ломаной) регрессии предположение о линейной корреляционной зависимости между переменными.

Для расчета R_{yx} по уравнению регрессии $y_x = 0{,}6762x - 4{,}79$ (см. пример **12.1**) находим значения y_{x_i}, представленные в предпоследней графе табл. 12.4. Затем аналогично $\delta_{-}^2 = 502{,}0/50 = 10{,}04$ и $R_{yx}=\sqrt{\dfrac{10{,}04}{18{,}23}}=\sqrt{0{,}551}=0{,}742$. Как и следовало ожидать, R_{yx} оказался равным r (небольшое расхождение объясняется округлением промежуточных результатов при вычислении R_{yx}). Поэтому в случае линейной связи нет смысла вычислять R_{yx}, а достаточно ограничиться вычислением r. Величина коэффициента детерминации $R_{yx}^2 = 0{,}551$ показывает, что вариация зависимой переменной Y (суточной выработки продукции) на

55,1% объясняется вариацией независимой переменной X (величиной основных производственных фондов).

Для проверки значимости η_{yx}, учитывая, что количество интервалов по группировочному признаку $m = 5$, по **(12.63)** найдем

$$F = \frac{0{,}754^2(50-5)}{(1-0{,}754)^2(5-1)} = 14{,}82.$$

Табличное значение $F_{0,05;4;45}=2{,}57$. Так как $F > F_{0,05;4;45}$, то η_{yx} значимо отличается от нуля. Аналогично проверяется значимость R_{yx}. По **(12.64)** $F = \frac{0{,}742^2(50-2)}{(1-0{,}742^2)} = 58{,}8$. Так как $F > F_{0,05;1;48} = 4{,}04$, то индекс корреляции R_{yx} значим. ▶

12.7. Понятие о многомерном корреляционном анализе. Множественный и частный коэффициенты корреляции

Экономические явления чаще всего адекватно описываются многофакторными моделями. Поэтому возникает необходимость обобщить рассмотренную выше двумерную корреляционную модель на случай нескольких переменных.

Пусть имеется совокупность случайных переменных $X_1, X_2, ..., X_i, ..., X_j, ..., X_p$, имеющих совместное нормальное распределение. В этом случае матрицу

$$Q_p = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}, \qquad \textbf{(12.65)}$$

составленную из парных коэффициентов корреляции ρ_{ij} ($i, j = 1, 2, ..., p$), определяемых по формуле **(9.2)**, будем называть *корреляционной. Основная задача многомерного корреляционного анализа состоит в оценке корреляционной матрицы Q_p по выборке.* Эта задача решается определением *матрицы выборочных коэффициентов корреляции*:

$$q_p = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}, \qquad \textbf{(12.66)}$$

где r_{ij} ($i, j = 1,2, ..., p$) определяется по формуле **(12.30)** или ее модификациям.

В многомерном корреляционном анализе рассматривают две типовые задачи:

а) определение тесноты связи одной из переменных с совокупностью остальных $(p - 1)$ переменных, включенных в анализ;

б) определение тесноты связи между переменными при фиксировании или исключении влияния остальных q переменных, где $q \le (p-2)$.

Эти задачи решаются с помощью множественных и частных коэффициентов корреляции.

Множественный коэффициент корреляции. Теснота линейной взаимосвязи одной переменной X_i с совокупностью других $(p - 1)$ переменных X_j, рассматриваемой в целом, измеряется с помощью *множественного* (или *совокупного*) *коэффициента корреляции* $\rho_{i.12...p}$, который является обобщением парного коэффициента корреляции ρ_{ij}. *Выборочный множественный*, или *совокупный, коэффициент корреляции* $R_{i.12...p}$, являющийся оценкой $R_{i.12...p}$, может быть вычислен по формуле:

$$R_{i.12...p} = \sqrt{1 - \frac{|q_p|}{q_{ii}}}, \qquad \textbf{(12.67)}$$

где $|q_p|$ — определитель матрицы q_p;

q_{ii} — алгебраическое дополнение элемента r_{ii} той же матрицы (равного 1).

В частности, в случае трех переменных ($p = 3$) из **(12.67)** следует, что

$$R_{i.jk} = \sqrt{\frac{r_{ij}^2 + r_{ik}^2 - 2r_{ij} \cdot r_{ik} \cdot r_{jk}}{1 - r_{jk}^2}}. \qquad \textbf{(12.68)}$$

Множественный коэффициент корреляции заключен в пределах $0 \le R \le 1$. Он не меньше, чем абсолютная величина любого парного или частного коэффициента корреляции с таким же первичным индексом.

С помощью множественного коэффициента корреляции (по мере приближения R к 1) делается *вывод о тесноте взаимосвязи, но не о ее направлении.* Величина R^2, называемая *выборочным множественным* (или *совокупным*) *коэффициентом детерминации*, показывает, какую долю вариации исследуемой переменной объясняет вариация остальных переменных.

Можно показать, что множественный коэффициент корреляции значимо отличается от нуля, если значение статистики

$$F = \frac{R^2(n-p)}{(1-R^2)(p-1)} > F_{\alpha;k_1,k_2}\ , \qquad \textbf{(12.69)}$$

где F_{α,k_1,k_2} — табличное значение F-критерия на уровне значимости α при числе степеней свободы $k_1 = p - 1$ и $k_2 = n - p$.

Частный коэффициент корреляции. Если переменные коррелируют друг с другом, то на величине парного коэффициента корреляции частично сказывается влияние других переменных. В связи с этим часто возникает необходимость исследовать *частную корреляцию* между переменными при исключении (элиминировании) влияния одной или нескольких других переменных.

Выборочным частным коэффициентом корреляции между переменными X_i и X_j при фиксированных значениях остальных $(p-2)$ переменных называется выражение

$$r_{ij.12\ldots p} = \frac{-q_{ij}}{\sqrt{q_{ii}q_{jj}}}, \qquad \textbf{(12.70)}$$

где q_{ij} и q_{jj} — алгебраические дополнения элементов r_{ij} и r_{jj} матрицы q_p. В частности, в случае трех переменных (p=3) из **(12.70)** следует, что

$$r_{ij.k} = \frac{r_{ij} - r_{ik}\cdot r_{jk}}{\sqrt{(1-r_{ik}^2)(1-r_{jk}^2)}}. \qquad \textbf{(12.71)}$$

Частный коэффициент корреляции $r_{ij.12\ldots p}$, как и парный коэффициент корреляции r, может принимать значения от -1 до 1. Кроме того, $r_{ij.12\ldots p}$, вычисленный на основе выборки объема

n, имеет такое же распределение, что и r, вычисленный по $(n-p+2)$ наблюдениям. Поэтому значимость частного коэффициента корреляции $r_{ij.12\ldots p}$ оценивают так же, как и коэффициента корреляции r (см. § 12.5), но при этом полагают $n' = n - p + 2$.

▷ **Пример 12.9.** Для исследования зависимости между производительностью труда (X_1), возрастом (X_2) и производственным стажем (X_3) была произведена выборка из 100 рабочих одной и той же специальности. Вычисленные парные коэффициенты корреляции оказались значимыми и составили: r_{12}=0,20; r_{13}=0,41; r_{23}=0,82. Вычислить множественный коэффициент корреляции $R_{1.23}$, частные коэффициенты корреляции и оценить их значимость на уровне $\alpha = 0{,}05$.

Р е ш е н и е. По **(12.68)** вычислим множественный коэффициент корреляции:

$$R_{1\,23} = \sqrt{\frac{0{,}20^2 + 0{,}41^2 - 2\cdot 0{,}20\cdot 0{,}41\cdot 0{,}82}{1-0{,}82^2}} = \sqrt{0{,}225} = 0{,}47\,,$$

т.е. между производительностью труда, с одной стороны, и возрастом и производственным стажем рабочих — с другой, существует заметная связь. Множественный коэффициент детерминации $R^2_{1.23} = 0{,}225$ показывает, что вариация производительности труда рабочих на 22,5% объясняется вариацией их возраста и производственного стажа.

Для оценки значимости $R_{1.23}$ по **(12.69)** вычислим

$$F = \frac{0{,}47^2\cdot(100-3)}{(1-0{,}47^2)\cdot(3-1)} = 14{,}1$$

и по таблицам F-распределения найдем $F_{0{,}05;2;97}$=3,09. Так как $F > F_{0{,}05;2;97}$, то $R_{1.23}$ значимо отличается от нуля.

По **(12.71)** вычислим частные коэффициенты корреляции:

$$r_{12.3} = \frac{0{,}20^2 - 0{,}41\cdot 0{,}82}{\sqrt{(1-0{,}41^2)(1-0{,}82^2)}} = -0{,}26$$

и аналогично $r_{13.2} = 0{,}44$; $r_{23.1} = 0{,}83$.

Оценим значимость $r_{12.3}$. Полагаем условно $n' = n - p + 2 = 100 - 3 + 2 = 99$. Статистика критерия по **(12.43)**:

$$t = \frac{-0{,}26 \cdot \sqrt{99-2}}{\sqrt{1-0{,}26^2}} = -2{,}65.$$

По таблице t-распределения Стьюдента находим $t_{0{,}05;97} = 1{,}99$. Так как $|t| > t_{0{,}95;97}$, то частный коэффициент корреляции $r_{12.3}$ значим. Тем более будут значимы бóльшие коэффициенты $r_{13.2}$ и $r_{23.1}$ (в этом можно убедиться таким же образом). ▶

Сравнивая частные коэффициенты корреляции $r_{ij.k}$ с соответствующими парными коэффициентами r_{ij}, видим, что за счет «очищения связи» наибольшему изменению подвергся коэффициент корреляции между производительностью труда (X_1) и возрастом (X_2) рабочих (изменилась не только его величина, но даже и знак: r_{12}=0,20; $r_{12.3}$=– 0,26, причем оба эти коэффициента значимы).

Итак, между производительностью труда (X_1) и возрастом (X_2) рабочих существует прямая корреляционная связь (r_{12}=0,20). Если же устранить (элиминировать) влияние переменной «производственный стаж» (X_3), то в чистом виде производительность труда (X_1) находится в обратной по направлению (и опять же слабой по тесноте) связи с возрастом рабочих (X_2) ($r_{12.3}$=– 0,26). Это вполне объяснимо, если рассматривать возраст только как показатель работоспособности организма на определенном этапе его жизнедеятельности. Подобным образом могут быть интерпретированы и другие частные коэффициенты корреляции.

Заканчивая краткое изложение корреляционного анализа количественных признаков, остановимся на двух моментах.

1. Задача научного исследования состоит в отыскании *причинных зависимостей*. Только знание истинных причин явлений позволяет правильно истолковывать наблюдаемые закономерности. Однако *корреляция как формальное статистическое понятие сама по себе не вскрывает причинного характера связи*. С помощью корреляционного анализа нельзя указать, какую переменную принимать в качестве причины, а какую — в качестве следствия. Например, рассматривая корреляционную связь между суточной выработкой продукции и величиной основных производственных фондов (см. пример **12.1**), изменение последней можно считать одной из причин изменения суточной выработки. Но, с другой стороны, необходимость повышения суточной

выработки продукции может повлечь за собой увеличение размера основных производственных фондов. Между урожайностью сельскохозяйственных культур и погодными условиями (температурой, количеством осадков и т.п.) существует корреляционная связь. Но здесь не возникает сомнений, какая переменная является следствием, а какая — причиной.

Иногда при наличии корреляционной связи ни одна из переменных не может рассматриваться причиной другой (например, зависимость между весом и ростом человека). Наконец, возможна *ложная корреляция* (*нонсенс-корреляция*), т.е. чисто формальная связь между переменными, не находящая никакого объяснения и основанная лишь на количественном соотношении между ними (таких примеров в статистической литературе приводится немало). Поэтому *при логических переходах от корреляционной связи между переменными к их причинной взаимообусловленности необходимо глубокое проникновение в сущность анализируемых явлений.*

2. *Не существует общеупотребительного критерия проверки определяющего требования корреляционного анализа — нормальности многомерного распределения переменных.* Учитывая свойства теоретической модели, обычно полагают, что отнесение к совместному нормальному закону возможно, если частные одномерные распределения переменных не противоречат нормальным распределениям (в этом можно убедиться, например, с помощью критериев согласия); если совокупность точек корреляционного поля частных двумерных распределений имеет вид более или менее вытянутого «облака» с выраженной линейной тенденцией.

Для проверки линейности связи пары признаков можно использовать расхождение между квадратами эмпирического корреляционного отношения η^2 и коэффициента корреляции r^2, учитывая, что статистика

$$F = \frac{(\eta^2 - r^2)(n-m)}{(m-2)(1-\eta^2)} \qquad \textbf{(12.72)}$$

(n — число наблюдений, m — число группировочных интервалов) имеет F-распределение с $k_1 = m - 2$ и $k_2 = n - m$ степенями свободы

▷ **Пример 12.10.** По данным табл. 12.1 на уровне значимости 0,05 проверить гипотезу о линейности корреляционной зависимости между переменными Y и X.

Р е ш е н и е. Имеем $n = 50$, $m = 5$. В примере **12.3** было получено $r = 0{,}740$, а в примере **12.7** — $\eta = 0{,}754$. По формуле **(12.72)**

$$F = \frac{(0{,}754^2 - 0{,}740^2)(50-5)}{(5-2)(1-0{,}754^2)} = 0{,}727 .$$

Так как $F < F_{0,05;3;45}$=2,82 (см. табл. VI приложений), то гипотеза о линейности корреляционной зависимости между Y и X не отвергается. ▶

Многомерный корреляционный анализ позволяет с помощью корреляционной матрицы **(12.66)** получить оценку модельного уравнения регрессии — линейного уравнения множественной регрессии. Однако это проще сделать с помощью регрессионного анализа (см. гл. 13).

12.8. Ранговая корреляция

До сих пор мы анализировали зависимости между *количественными* переменными, измеренными в так называемых *количественных* шкалах, т.е. в шкалах с непрерывным множеством значений, позволяющих выявить, на сколько (или во сколько раз) проявление признака у одного объекта больше (меньше), чем у другого (например, производительность труда, себестоимость продукции и т.п.).

Вместе с тем на практике часто встречаются с необходимостью изучения связи между *ординальными* (*порядковыми*) переменными, измеренными в так называемой *порядковой* шкале. В этой шкале можно установить лишь п о р я д о к, в котором объекты выстраиваются по степени проявления признака (например, качество жилищных условий, тестовые баллы, экзаменационные оценки и т.п.). Если, скажем, по некоторой дисциплине два студента имеют оценки «отлично» и «удовлетворительно», то можно лишь утверждать, что уровень подготовки по этой дисциплине первого студента выше (больше), чем второго, но нельзя сказать, на сколько или во сколько раз больше.

Оказывается, что в таких случаях проблема оценки тесноты связи разрешима, если упорядочить, или ранжировать, объекты анализа по степени выраженности измеряемых признаков. При этом каждому объекту присваивается определенный номер, называемый *рангом*. Например, объекту с наименьшим проявлением (значением) признака присваивается ранг 1, следующему за ним — ранг 2 и т.д. Объекты можно располагать и в порядке убывания проявления (значений) признака. Если объекты ранжированы по двум признакам, то имеется возможность оценить тесноту связи между признаками, основываясь на рангах, т.е. тесноту *ранговой корреляции*.

Коэффициент ранговой корреляции Спирмена находится по формуле:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(r_i - s_i)^2}{n^3 - n}, \qquad \textbf{(12.73)}$$

где r_i и s_i — ранги i-го объекта по переменным X и Y, n — число пар наблюдений.

Если ранги всех объектов равны ($r_i = s_i$, $i = 1,2,....,n$), то $\rho = 1$, т.е. при полной прямой связи $\rho = 1$. При полной обратной связи, когда ранги объектов по двум переменным расположены в обратном порядке, можно показать, что $\sum_{i=1}^{n}(r_i - s_i)^2 = (n^3 - n)/3$ и по формуле **(12.72)** $\rho = -1$. Во всех остальных случаях $|\rho| < 1$.

При ранжировании иногда сталкиваются со случаями, когда невозможно найти существенные различия между объектами по величине проявления рассматриваемого признака. Объекты, как говорят, оказываются *связанными*. Связанным объектам приписывают одинаковые средние ранги, такие, чтобы сумма всех рангов оставалась такой же, как и при отсутствии связанных рангов. Например, если четыре объекта оказались равнозначными в отношении рассматриваемого признака и невозможно определить, какие из четырех рангов (4,5,6,7) приписать этим объектам, то каждому объекту приписывается средний ранг, равный (4+5+6+7)/4= 5,5.

При наличии *связанных рангов* ранговый коэффициент корреляции Спирмена вычисляется по формуле:

$$\rho = 1 - \frac{\sum_{i=1}^{n}(r_i - s_i)^2}{\frac{1}{6}(n^3 - n) - (T_r + T_s)}, \qquad \textbf{(12.74)}$$

где $T_r = \frac{1}{12}\sum_{i=1}^{m_r}(t_r^3 - t_r)$; $\qquad T_s = \frac{1}{12}\sum_{i=1}^{m_s}(t_s^3 - t_s)$; $\qquad$ **(12.75)**

m_r, m_s — число групп неразличимых рангов у переменных X и Y;

t_r, t_s — число рангов, входящих в группу неразличимых рангов переменных X и Y.

При проверке значимости ρ исходят из того, что в случае справедливости нулевой гипотезы об отсутствии корреляционной связи между переменными при $n > 10$ статистика

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}} \qquad \textbf{(12.76)}$$

имеет t-распределение Стьюдента с $k = n - 2$ степенями свободы. Поэтому ρ значим на уровне α, если фактически наблюдаемое значение t будет больше критического (по абсолютной величине), т.е. $|t| > t_{1-\alpha,n-2}$, где $t_{1-\alpha,n-2}$ — табличное значение t-критерия Стьюдента, определенное на уровне значимости α при числе степеней свободы $k = n - 2$.

▷ **Пример 12.11.** По результатам тестирования 10 студентов по двум дисциплинам A и B на основе набранных баллов получены следующие ранги (табл. 12.5). Вычислить ранговый коэффициент корреляции Спирмена и проверить его значимость на уровне $\alpha = 0{,}05$.

Р е ш е н и е. Разности рангов и их квадраты поместим в последних двух строках табл. 12.5.

Таблица 12.5

Ранги по дисциплинам		*Студент, i*										*Всего*
		1	2	3	4	5	6	7	8	9	10	
A	r_i	2	4	5	1	7,5	7,5	7,5	7,5	3	10	55
B	s_i	2,5	6	4	1	2,5	7	8	9,5	5	9,5	55
$r_i - s_i$		−0,5	−2	1	0	5	0,5	−0,5	−2	−2	0,5	—
$(r_i - s_i)^2$		0,25	4	1	0	25	0,25	0,25	4	4	0,25	39

По формуле **(12.73)** $\rho = 1 - \dfrac{6 \cdot 39}{10^3 - 10} = 0{,}763$. Однако формула **(12.73)** не учитывает наличия связанных рангов.

По дисциплине A имеем $m_r = 1$ — одну группу неразличимых рангов с t_r=4 рангами; по дисциплине B — $m_s = 2$ — две группы неразличимых рангов по $t_s = 2$ ранга. Поэтому по формуле **(12.75)**

$$T_r = \frac{1}{12}\left(4^3 - 4\right) = 5, \quad T_s = \frac{1}{12}\left[\left(2^3 - 2\right) + \left(2^3 - 2\right)\right] = 1.$$

Находим по формуле **(12.74)**

$$\rho = 1 - \frac{39}{\frac{1}{6}\left(10^3 - 10\right) - \left(5 + 1\right)} = 0{,}755.$$

Для проверки значимости ρ по формуле **(12.76)**[1] вычислим $t = 0{,}755\dfrac{\sqrt{10-2}}{\sqrt{1-0{,}755^2}} = 3{,}26$ и найдем по табл. IV приложений $t_{0,95;8}$=2,31. Так как $t > t_{0,95;8}$, то ранговый коэффициент корреляции ρ значим на 5%-ном уровне. Связь между оценками двух дисциплин достаточно тесная. ◀

Коэффициент ранговой корреляции Кендалла находится по формуле:

$$\tau = 1 - \frac{4K}{n(n-1)}, \qquad \textbf{(12.77)}$$

где K — статистика Кендалла[2].

Для определения K необходимо ранжировать объекты по одной переменной в порядке возрастания рангов (1, 2,...,n) и определить соответствующие их ранги ($r_1, r_2, ..., r_n$) по другой переменной. Статистика K равна общему числу *инверсий* (нарушений порядка, когда большее число стоит слева от меньшего) в ранговой последовательности (*ранжировке*) $r_1, r_2, ..., r_n$. При полном совпадении двух ранжировок имеем $K = 0$ и $\tau = 1$; при полной противоположности можно показать, что $K = n(n-1)/2$ и $\tau = -1$. Во всех остальных случаях $|\tau| < 1$.

При проверке значимости τ исходят из того, что в случае справедливости нулевой гипотезы об отсутствии корреляционной связи между переменными (при $n > 10$) τ имеет приближенно нормальный закон распределения с математическим ожиданием, равным нулю, и средним квадратическим отклоне-

[1] В примерах **12.11** и **12.12** использованы приближенно при n=10 критерии проверки значимости соответственно ρ и τ, справедливые, вообще говоря, при n>10.

[2] Формула для расчета τ при наличии связанных рангов здесь не приводится.

нием $s_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}}$. Поэтому τ значим на уровне α, если значение статистики

$$t = \frac{\tau - 0}{s_\tau} = \tau\sqrt{\frac{9n(n-1)}{2(2n+5)}} \qquad \textbf{(12.78)}$$

больше критического $t_{1-\alpha}$, где $\Phi(t_{1-\alpha}) = 1 - \alpha$.

Поясним вычисление рангового коэффициента корреляции Кендалла на примере.

▷ **Пример 12.12.** В результате анкетного обследования для 10 важнейших видов оборудования, используемого судоводителями во время вахты, получены следующие ранги по важности оборудования X и по частоте его использования Y (см. табл. 12.6). Вычислить ранговый коэффициент Кендалла и оценить его значимость на уровне $\alpha = 0{,}05$.

Р е ш е н и е. В последней строке табл. 12.6 представлены значения числа инверсий в ранжировках по переменной Y для различных рангов по переменной X.

Таблица 12.6

Ранг	*Тип оборудования*										*Всего*
	А	*Б*	*В*	*Г*	*Д*	*Е*	*Ж*	*З*	*И*	*К*	
Важность оборудования X, n	1	2	3	4	5	6	7	8	9	10	—
Частота использования Y, r_i	1	4	2	6	3	9	10	8	7	5	—
Число инверсий	0	2	0	2	0	3	3	2	1	0	$K = 13$

Найдем, например, число инверсий при ранге $n = 6$ по переменной X. Тогда соответствующий ранг по переменной Y $r_6 = 9$ и с учетом последующих рангов (см. табл. 12.6) имеем ранжировку по Y (9, 10, 8, 7, 5).

Из пар чисел (перестановок) (9, 10), (9, 8), (9, 7), (9, 5) инверсии (нарушения порядка, когда большее число стоит слева от меньшего) имеются у трех последних пар, т.е. число инверсий равно 3. Аналогично определяются и другие значения числа инверсий и находится их сумма $K = 13$. Теперь по формуле **(12.77)**

$$\tau = 1 - \frac{4 \cdot 13}{10 \cdot 9} = 0{,}422.$$

Оценим значимость τ. Вычислим по формуле **(12.78)** значение статистики $t = 0{,}422\sqrt{\frac{9 \cdot 10(10-1)}{2(2 \cdot 10 + 5)}} = 8{,}49$, по табл. IV приложений $t_{0{,}95}$=1,96. Так как $t > t_{0{,}95}$, то ранговый коэффициент корреляции Кендалла значим на 5%-ном уровне. Связь между рассматриваемыми переменными умеренная. ▶

Сравнивая коэффициенты ранговой корреляции ρ (Спирмена) и τ (Кендалла), можно отметить, что хотя вычисление τ более трудоемко, коэффициент τ обладает некоторыми преимуществами перед ρ при исследовании его статистических свойств (например, возможностью приближенного построения доверительного интервала для τ) и большим удобством его пересчета при добавлении к n статистически обследованным объектам новых, т.е. при удлинении анализируемых ранжировок.

Значения коэффициентов ρ и τ тесно связаны между собой.

При умеренно больших значениях n ($n > 10$) и при условии, что абсолютные величины значений этих коэффициентов не слишком близки к единице, их связывает простое приближенное соотношение $\rho \approx 1{,}5\tau$.

Ранговые коэффициенты корреляции ρ и τ могут быть использованы и для оценки тесноты связи между обычными количественными переменными, измеряемыми в интервальных шкалах. Достоинство ρ и τ здесь заключается в том, что нахождение этих коэффициентов не требует нормального распределения переменных, линейной связи между ними (хотя и предполагает монотонность функции регрессии, отражающей эту связь). Однако необходимо учитывать, что при переходе от первоначальных значений переменных к их рангам происходит определенная потеря информации. Чем теснее связь, чем меньше корреляционная зависимость между переменными отличается от линейной, тем ближе коэффициент Спирмена ρ к коэффициенту парной корреляции r.

В практике статистических исследований встречаются случаи, когда совокупность объектов характеризуется не двумя, а несколькими последовательностями рангов (ранжировками) и

необходимо установить статистическую связь между несколькими переменными. Такие задачи возникают, например, при анализе экспертных оценок, когда необходимо установить меру их согласованности.

В качестве такого измерителя используют ***коэффициент конкордации (согласованности) рангов Кендалла*** W, определяемый по формуле[1]:

$$W = \frac{12\sum_{i=1}^{n} D^2}{m^2(n^3 - n)}, \qquad (12.79)$$

где n — число объектов,

m — число анализируемых порядковых переменных,

$$D = \sum_{j=1}^{m} r_{ij} - \frac{m(n+1)}{2} \qquad (12.80)$$

— отклонение суммы рангов объекта от средней их суммы для всех объектов, равной $m(n+1)/2$.

Можно доказать, что значения коэффициента W заключены на отрезке [0; 1], т.е. $0 \le W \le 1$, причем $W = 1$ при совпадении всех ранжировок.

Проверка значимости коэффициента конкордации W основана на том, что в случае справедливости нулевой гипотезы об отсутствии корреляционной связи при $n > 7$ статистика $m(n-1)W$ имеет приближенно χ^2-распределение с $k = n - 1$ степенями свободы. Поэтому W значим на уровне α, если

$$m(n-1)W > \chi^2_{\alpha,n-1}. \qquad (12.81)$$

▷ **Пример 12.13.** Группа из 5 экспертов оценивает качество изделий, изготовленных на 7 предприятиях. Их предпочтения представлены в табл. 12.7. Вычислить коэффициент конкордации рангов и оценить его значимость на уровне $\alpha = 0{,}05$.

Р е ш е н и е. В итоговой строке табл. 12.7 приведены суммы рангов изделий по каждому из 7 предприятий, полученных от 5 экспертов. Общая сумма рангов равна 140. Средняя сумма рангов равна $m(n+1)/2 = 5(7+1)/2=20$ или, иначе, $140/7=20$.

[1] Формула для расчета W при наличии связанных рангов здесь не приводится.

Таблица 12.7

Эксперт, j	*Предприятие, i*							*Итого*
	1	2	3	4	5	6	7	
1	1	3	4	2	6	7	5	
2	1	2	5	3	6	4	7	
3	2	1	7	5	6	4	3	
4	1	2	4	6	3	5	7	
5	3	1	5	4	2	6	7	
Сумма рангов $\sum_{j=1}^{5} r_{ij}$	8	9	25	20	23	26	29	140
D	−12	−11	5	0	3	6	9	–
D^2	144	121	25	0	9	36	81	416

В предпоследней строке табл. 12.7 помещены разности $D = \sum_{j=1}^{5} r_{ij} - 20$, а в последней строке — их квадраты D^2.

Коэффициент конкордации по формуле **(12.79)** $W = \frac{12 \cdot 416}{5^2(7^3 - 7)} = 0{,}594$. Оценим значимость W[1]. Вычислим $m(n-1)W = 5 \cdot 6 \cdot 0{,}594 = 17{,}83$; по табл. V приложений $\chi^2_{0,05;6} = 12{,}59$. Так как $m(n-1)W > \chi^2_{0,05;6}$, то коэффициент конкордации W значим на 5%-ном уровне. Таким образом, существует достаточно тесная согласованность мнений экспертов. ▶

Корреляционный анализ может быть использован и при оценке взаимосвязи *качественных* (*категоризованных*) признаков (переменных), представленных в так называемой *номинальной* шкале, в которой возможно лишь различение объектов по возможным состояниям, градациям (например, пол, социальное положение, профессия и т.п.). Здесь в качестве соответствующих показателей могут быть использованы коэффициенты *ассоциации*, *контингенции* (*сопряженности*), *бисериальной корреляции*. Эти вопросы рассмотрены, например, в [2], [32].

[1] Используем приближенно при $n=7$ критерий проверки значимости W, справедливый, вообще говоря, при $n>7$.

Упражнения

12.14. Распределение 60 предприятий химической промышленности по энерговооруженности труда Y (кВт · ч) и фондовооруженности X (млн руб.) дано в таблице.

y \ x	0—4,5	4,5—9,0	9,0—13,5	13,5—18,0	18,0—22,5	Итого
0—1,4	4	1	—	—	—	5
1,4—2,8	4	2	—	—	—	6
2,8—4,2	2	8	1	—	—	11
4,2—5,6	—	1	20	4	—	25
5,6—7,0	—	—	3	3	3	9
7,0—8,4	—	—	—	1	3	4
Итого	10	12	24	8	6	60

Необходимо: а) найти групповые средние x_j и y_i и построить эмпирические линии регрессии; б) оценить тесноту и направление связи между переменными с помощью коэффициента корреляции; проверить значимость коэффициента корреляции и построить для него 95%-ный доверительный интервал; в) вычислить эмпирические корреляционные отношения и оценить их значимость на 5%-ном уровне; г) на уровне значимости 0,05 проверить гипотезу о линейной корреляционной зависимости между переменными Y и X; д) найти уравнения прямых регрессии, построить их графики и найти 95%-ные доверительные интервалы для коэффициентов регрессии.

12.15. Имеются следующие данные об уровне механизации работ X(%) и производительности труда Y (т/ч) для 14 однотипных предприятий:

x_i	32	30	36	40	41	47	56	54	60	55	61	67	69	76
y_i	20	24	28	30	31	33	34	37	38	40	41	43	45	48

Необходимо: а) оценить тесноту и направление связи между переменными с помощью коэффициента корреляции; проверить значимость коэффициента корреляции и построить для него 95%-ный доверительный интервал; б) найти уравнения прямых регрессии.

12.16. При исследовании корреляционной зависимости по данным 20 предприятий между капиталовложениями X (млн руб.) и выпуском продукции Y (млн руб.) получены следующие уравнения регрессии: $y=1{,}2x+2$ и $x=0{,}7y+2$. Найти: а) коэффициент корреляции между рассматриваемыми признаками и оценить его значимость на 5%-ном уровне; б) средние значения капиталовложений и выпуска продукции. Согласуется ли полученный в п. а) результат с утверждением о том, что генеральный коэффициент корреляции между X и Y равен 0,95?

12.17. При исследовании корреляционной зависимости между ценой на нефть X и индексом нефтяных компаний Y получены следующие данные: $x=16{,}2$ (ден. ед.), $y=4000$ (усл. ед.), $s_x^2=4$, $s_y^2=500$, $\mu=40$. Необходимо: а) составить уравнения регрессии Y по X и X по Y; б) используя соответствующее уравнение регрессии, найти среднюю величину индекса при цене на нефть 16,5 ден. ед.

12.18. При исследовании корреляционной зависимости между объемом продукции X (единиц) и ее себестоимости Y (тыс. руб.) получено следующее уравнение регрессии Y по X: $y_x=-0{,}0004x+4{,}22$. Составить уравнение регрессии X по Y, если коэффициент корреляции между этими признаками оказался равным $-0{,}8$, а средний объем продукции $x=3000$ единиц.

12.19. С целью исследования влияния факторов X_1 — среднемесячного количества профилактических наладок автоматической линии и X_2 — среднемесячного числа обрывов нити на показатель Y — среднемесячную характеристику качества ткани (в баллах) по данным 37 предприятий легкой промышленности были вычислены парные коэффициенты корреляции: $r_{y_1}=0{,}105$, $r_{y_2}=0{,}024$ и $r_{12}=0{,}996$. Определить: а) частные коэффициенты корреляции $r_{y1.2}$ и $r_{y2.1}$ и оценить их значимость на 5%-ном уровне; б) множественный коэффициент корреляции $R_{y.12}$ и оценить его значимость на уровне $\alpha=0{,}05$; в) множественный коэффициент детерминации. Пояснить смысл полученных коэффициентов.

12.20. При приеме на работу семи кандидатам на вакантные должности было предложено два теста. Результаты тестирования (в баллах) приведены в таблице:

Тест	*Кандидат*						
	1	2	3	4	5	6	7
1	31	82	25	26	53	30	29
2	21	55	8	27	32	42	26

Вычислить ранговые коэффициенты корреляции Спирмена и Кендалла между результатами тестирования по двум тестам и на уровне $\alpha = 0{,}05$ оценить их значимость.

12.21. На соревнованиях по фигурному катанию девять судей выставили следующие балльные оценки 10 фигуристам:

Фигурист	*Судья*								
	1	2	3	4	5	6	7	8	9
1	6,0	5,8	5,7	5,8	6,0	5,9	5,9	5,9	5,8
2	5,4	5,3	5,2	5,3	5,4	5,5	5,6	5,3	5,1
3	5,2	5,0	4,9	5,1	5,2	5,0	4,8	5,3	4,9
4	5,9	5,9	5,8	5,7	5,9	5,8	6,0	5,8	5,7
5	5,0	4,9	4,9	4,9	5,1	5,0	5,0	4,8	4,7
6	5,6	5,5	5,4	5,4	5,5	5,5	5,7	5,6	5,5
7	4,8	4,7	4,6	4,6	4,8	4,9	5,0	4,6	4,5
8	5,4	5,6	5,4	5,5	5,6	5,7	5,4	5,3	5,2
9	5,8	5,7	5,6	5,7	5,8	5,9	5,6	5,7	5,8
10	5,3	5,2	5,1	5,4	5,5	5,4	5,2	5,3	5,2

Вычислить коэффициент конкордации рангов и оценить его значимость на уровне $\alpha = 0{,}05$.

Глава 13 Регрессионный анализ

В практике экономических исследований очень часто имеющиеся данные нельзя считать выборкой из многомерной нормальной совокупности, например, когда одна из рассматриваемых переменных не является случайной или когда линия регрессии явно не прямая и т.п. В этих случаях пытаются определить кривую (поверхность), которая дает наилучшее (в смысле метода наименьших квадратов) приближение к исходным данным. Соответствующие методы приближения получили название *регрессионного анализа.*

Задачами регрессионного анализа *являются установление формы зависимости между переменными, оценка функции регрессии, оценка неизвестных значений (прогноз значений) зависимой переменной.*

13.1. Основные положения регрессионного анализа. Парная регрессионная модель

В регрессионном анализе рассматривается *односторонняя зависимость случайной зависимой переменной Y от одной (или нескольких) неслучайной независимой переменной X,* называемой часто *объясняющей переменной*[1]. Такая зависимость может возникнуть, например, в случае, когда при каждом фиксированном значении X соответствующие значения Y подвержены случайному разбросу за счет действия неконтролируемых факторов. Указанная зависимость Y от X (иногда ее называют ***регрессионной***) может быть представлена также в виде модельного уравнения регрессии **(12.1)**. В силу воздействия неучтенных случайных факторов и причин отдельные наблюдения y будут в большей или меньшей мере отклоняться от функции регрессии $\varphi(x)$. В этом случае *уравнение взаимосвязи двух переменных (парная регрессионная модель)* может быть представлено в виде:

$$Y = \varphi(X) + \varepsilon,$$

где ε — случайная переменная, характеризующая отклонение от функции регрессии. Эту переменную будем называть *возмущаю-*

[1] В литературе Y называют также *функцией отклика, объясняемой, выходной, результирующей, эндогенной переменной, результативным признаком,* а X — *входной, предсказывающей, предикторной, экзогенной переменной, фактором; регрессором, факторным признаком.*

щей или просто *возмущением*[1]. Таким образом, *в регрессионной модели зависимая переменная Y есть некоторая функция* $\varphi(X)$ *с точностью до случайного возмущения* ε.

Рассмотрим *линейный регрессионный анализ*, для которого функция $\varphi(X)$ л и н е й н а относительно оцениваемых параметров:

$$M_x(Y)=\beta_0+\beta_1 x. \qquad \textbf{(13.1)}$$

Предположим, что для оценки параметров линейной функции регрессии **(13.1)** взята выборка, содержащая n пар значений переменных (x_i, n_i), где $i=1,2,\ldots, n$. В этом случае *линейная парная регрессионная модель* имеет вид:

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i. \qquad \textbf{(13.2)}$$

Отметим **основные предпосылки регрессионного анализа**:

1. В модели **(13.2)** *возмущение*[2] ε_i *(или зависимая переменная* y_i*) есть величина случайная, а объясняющая переменная* x_i *— величина неслучайная*[3].

2. *Математическое ожидание возмущения* ε_i *равно нулю*:

$$M(\varepsilon_i)=0. \qquad \textbf{(13.3)}$$

(или математическое ожидание зависимой переменной y_i равно линейной функции регрессии: $M(y_i)=\beta_0+\beta_1 x_i$).

3. *Дисперсия возмущения* ε_i *(или зависимой переменной* y_i*) постоянна для любого* i:

$$D(\varepsilon_i)=\sigma^2 \qquad \textbf{(13.4)}$$

(или $D(y_i)=\sigma^2$ — условие *гомоскедастичности* или *равноизменчивости* возмущения (зависимой переменной)).

4. *Возмущения* ε_i *и* ε_j *(или переменные* y_i *и* y_j*) не коррелированы*[4]:

$$M(\varepsilon_i \varepsilon_j)=0 \quad (i \neq j). \qquad \textbf{(13.5)}$$

5. *Возмущение* ε_i *(или зависимая переменная* y_i*) есть нормально распределенная случайная величина.*

---

[1] Переменную ε называют также *остаточной* или *остатком*, либо *ошибкой*.

[2] Во всех предпосылках $i=1,2,\ldots,n$.

[3] При этом предполагается, что среди значений x_i $(i=1,2,\ldots,n)$, по крайней мере, не все одинаковые, так что имеет смысл формула **(12.17)** для коэффициента регрессии.

[4] Требование некоррелированности $K_{\varepsilon_i\varepsilon_j}=0$ с учетом **(5.32)** и **(13.3)** приводит к условию **(13.5)**: $K_{\varepsilon_i\varepsilon_j}=M[(\varepsilon_i-0)(\varepsilon_j-0)]=M(\varepsilon_i\varepsilon_j)=0$. При выполнении предпосылки **5** это требование равносильно независимости переменных.

Для получения уравнения регрессии достаточно первых четырех предпосылок. Требование выполнения пятой предпосылки (т.е. рассмотрение «*нормальной регрессии*») необходимо для оценки т о ч н о с т и уравнения регрессии и его параметров.

Оценкой модели **(13.2)** по выборке является уравнение регрессии $y_x=b_0+b_1 x$ **(12.8)**. Параметры этого уравнения b_0 и b_1 определяются на основе метода наименьших квадратов. Об их нахождении подробно см. в § 12.2.

Воздействие неучтенных случайных факторов и ошибок наблюдений в модели **(13.2)** определяется с помощью *дисперсии возмущений* (*ошибок*) или *остаточной дисперсии* σ^2. Несмещенной оценкой этой дисперсии является *выборочная остаточная дисперсия*

$$s^2=\frac{\sum_{i=1}^{n}(y_{x_i}-y_i)^2}{n-2}=\frac{\sum_{i=1}^{n}e_i^2}{n-2}, \qquad \textbf{(13.6)}$$

где y_{x_i} — *групповая средняя*, найденная по уравнению регрессии;

$e_i=y_{x_i}-y_i$ — *выборочная оценка возмущения* ε_i или *остаток регрессии*[1].

В знаменателе выражения **(13.6)** стоит число степеней свободы $n-2$, а не n, так как две степени свободы теряются при определении двух параметров прямой b_0 и b_1.

13.2. Интервальная оценка функции регрессии

Построим **доверительный интервал для функции регрессии**, т.е. для условного математического ожидания $M_x(Y)$, который с заданной надежностью (доверительной вероятностью) $\gamma=1-\alpha$ накрывает неизвестное значение $M_x(Y)$.

Найдем дисперсию групповой средней y_x, представляющей выборочную оценку $M_x(Y)$. С этой целью уравнение регрессии **(12.15)** представим в виде:

$$y_x=\bar{y}+b_1(x-\bar{x}). \qquad \textbf{(13.7)}$$

На рис. 13.1 линия регрессии **(13.7)** изображена графически. Для произвольного наблюдаемого значения y_i выделены его составляющие: средняя $\bar{y}$, приращение $b_1(x_i-\bar{x})$, образующие расчетное значение y_{x_i}, и возмущение e_i.

---

[1] e_i называют также *невязкой*.

Дисперсия групповой средней равна сумме дисперсий двух независимых[1] слагаемых выражения (13.7):

$$\sigma^2_{y_x} = \sigma^2_{\bar{y}} + \sigma^2_{b_1}(x-\bar{x})^2. \qquad (13.8)$$

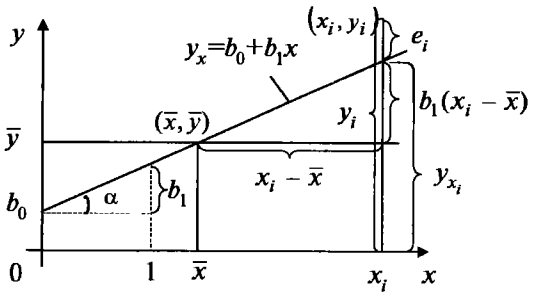


Рис. 13.1

(Здесь учтено, что $(x-\bar{x})$ — неслучайная величина, при вынесении которой за знак дисперсии ее необходимо возвести в квадрат.)

Дисперсия выборочной средней $\bar{y}$ согласно (9.16)

$$\sigma^2_{\bar{y}} = \frac{\sigma^2}{n}. \qquad (13.9)$$

Для нахождения дисперсии $\sigma^2_{b_1}$ представим коэффициент регрессии в виде[2]:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}. \qquad (13.10)$$

Тогда

$$\sigma^2_{b_1} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}. \qquad (13.11)$$

Найдем оценку дисперсии групповых средних (13.8), учитывая (13.9) и (13.11) и заменяя σ^2 ее оценкой s^2:

[1] Доказательство этого факта здесь не приводится.

[2] Раскрыв скобки и разделив числитель и знаменатель выражения (13.10) на n, нетрудно получить уже знакомое выражение (12.17).

$$s^2_{y_x} = s^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right). \qquad (13.12)$$

Исходя из того, что статистика $t = \dfrac{y_x - M_x(Y)}{s_{y_x}}$ имеет t-распределение Стьюдента с $k=n-2$ степенями свободы, можно (аналогично тому, как это сделано в § 9.7) построить **доверительный интервал для условного математического ожидания**

$$y_x - t_{1-\alpha;k}\cdot s_{y_x} \le M_x(Y) \le y_x + t_{1-\alpha;k}\cdot s_{y_x}, \qquad (13.13)$$

где $s_{y_x} = \sqrt{s^2_{y_x}}$ — стандартная ошибка групповой средней y_x.

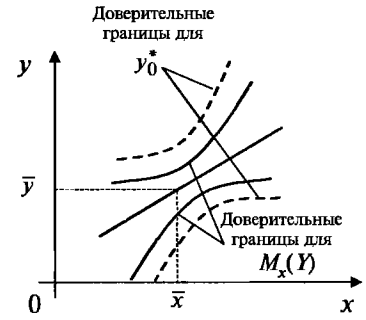


Рис. 13.2

Из формул (13.12) и (13.13) видно, что величина доверительного интервала зависит от значения объясняющей переменной x: при $x=\bar{x}$ она минимальна, а по мере удаления x от $\bar{x}$ величина доверительного интервала увеличивается (рис. 13.2). Таким образом, прогноз значений (определение неизвестных значений) зависимой переменной y по уравнению регрессии оправдан, если значение объясняющей переменной не выходит за диапазон ее значений по выборке (причем тем более точный, чем ближе x к $\bar{x}$). Другими словами, *экстраполяция кривой регрессии, т.е. ее использование вне пределов обследованного диапазона значений объясняющей переменной* (даже если она оправдана для рассматриваемой переменной исходя из смысла решаемой задачи) *может привести к значительным погрешностям.*

Построенная доверительная область для $M_x(Y)$ (см. рис. 13.2) определяет местоположение модельной линии регрессии (т.е. условного математического ожидания), но не отдельных возможных значений зависимой переменной, которые отклоняются от средней. Поэтому при определении **доверительного интервала**

для индивидуальных значений y_0^* зависимой переменной необходимо учитывать еще один источник вариации — *рассеяние вокруг линии регрессии*, т.е. в оценку суммарной дисперсии $s^2_{y_x}$ следует включить величину s^2. В результате оценка дисперсии индивидуальных значений y_0 при $x = x_0$ равна

$$s^2_{y_0} = s^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right], \qquad \textbf{(13.14)}$$

а соответствующий доверительный интервал для прогнозов индивидуальных значений y_0^* будет определяться по формуле:

$$y_{x_0} - t_{1-\alpha,n-2}s_{y_0} \le y_0^* \le y_{x_0} + t_{1-\alpha,n-2}s_{y_0}. \qquad \textbf{(13.15)}$$

▷ **Пример 13.1.** Имеются следующие данные (условные) о сменной добыче угля на одного рабочего Y (т) и мощности пласта X (м), характеризующие процесс добычи угля в 10 шахтах.

Таблица 13.1

i	1	2	3	4	5	6	7	8	9	10
x_i	8	11	12	9	8	8	9	9	8	12
y_i	5	10	10	7	5	6	6	5	6	8

Оценить сменную среднюю добычу угля на одного рабочего для шахт с мощностью пласта 8 м. Найти 95%-ные доверительные интервалы для индивидуального и среднего значений сменной добычи угля на 1 рабочего для таких же шахт.

Р е ш е н и е. Вначале аналогично тому, как это сделано в примере **12.1**, составим уравнение регрессии[1]. Получим (рекомендуем

[1] В расчетах полагаем $n_{ij}=n_i=n_j=1$, $j=i$, а $\sum_{i=1}^{l}\sum_{j=1}^{m}$ заменяем $\sum_{i=1}^{n}$, так как исходные данные не сгруппированы.

читателю получить самостоятельно) $\sum_{i=1}^{10} x_i = 94$, $\sum_{i=1}^{10} x_i^2 = 908$, $\sum_{i=1}^{10} y_i = 68$, $\sum_{i=1}^{10} y_i^2 = 496$, $\sum_{i=1}^{10} x_i y_i = 664$ и уравнение регрессии $y_x = -2{,}75 + 1{,}016x$, т.е. при увеличении мощности пласта X на 1 м добыча угля на одного рабочего Y увеличивается в среднем на 1,016 т (в усл. ед.).

Надо оценить условное математическое ожидание $M_{x=8}(Y)$. Выборочной оценкой $M_{x=8}(Y)$ является групповая средняя $y_{x=8}$, которую найдем по уравнению регрессии: $y_{x=8} = -2{,}75 + 1{,}016 \cdot 8 = 5{,}38$ (т).

Для построения доверительного интервала для $M_{x=8}(Y)$ необходимо знать дисперсию его оценки, т.е. $s^2_{y_{x=8}}$. Составим вспомогательную таблицу (табл. 13.2) с учетом того, что $\bar{x} = 9{,}4$ (м), а значения y_{x_i} определяются по полученному уравнению регрессии.

Таблица 13.2

x_i	8	11	12	9	8	8	9	9	8	12	Σ
$(x_i-\bar{x})^2$	1,96	2,56	6,76	0,16	1,96	1,96	0,16	0,16	1,96	6,76	24,40
$y_{x_i} = -2{,}75 + 1{,}016x_i$	5,38	8,43	9,44	6,39	5,38	5,38	6,39	6,39	5,38	9,44	—
$e_i^2 = (y_{x_i} - y_i)^2$	0,14	2,48	0,31	0,37	0,14	0,39	0,15	1,94	0,39	2,08	8,39

Теперь имеем по **(13.6)**: $s^2 = \frac{8{,}39}{10-2} = 1{,}049$, по **(13.12)**:

$$s^2_{y_{x=8}} = 1{,}049\left[\frac{1}{10}+\frac{(8-9{,}4)^2}{24{,}4}\right] = 0{,}189$$

и $$s_{y_{x=8}} = \sqrt{0{,}189} = 0{,}435 \text{ (т)}.$$

По табл. IV приложений $t_{0,95;8} = 2{,}31$. Теперь по **(13.13)** искомый доверительный интервал:

$$5{,}38 - 2{,}31 \cdot 0{,}435 \le M_{x=8}(Y) \le 5{,}38 + 2{,}31 \cdot 0{,}435$$

или $$4{,}38 \le M_{x=8}(Y) \le 6{,}38 \text{ (т)}.$$

Итак, средняя сменная добыча угля на одного рабочего для шахт с мощностью пласта 8 м с надежностью 0,95 находится в пределах от 4,38 до 6,38 т.

Чтобы построить доверительный интервал для индивидуального значения $y^*_{x_0=8}$, найдем дисперсию его оценки по **(13.14)**:

$$s^2_{y_{x_0=8}} = 1{,}049\left(1+\frac{1}{10}+\frac{(8-9{,}4)^2}{24{,}4}\right)=1{,}238$$

и $$s_{y_{x_0=8}} = \sqrt{1{,}238} = 1{,}113 \text{ (т)}.$$

Далее искомый доверительный интервал получим по **(13.15)**:

$$5{,}38-2{,}31\cdot 1{,}113 \le y^*_{x_0=8} \le 5{,}38+2{,}31\cdot 1{,}113$$

или $$2{,}81 \le y^*_{x_0=8} \le 7{,}95.$$

Таким образом, индивидуальная сменная добыча угля на одного рабочего для шахт с мощностью пласта 8 м с надежностью 0,95 находится в пределах от 2,81 до 7,95 т. ▶

13.3. Проверка значимости уравнения регрессии. Интервальная оценка параметров парной модели

Проверить **значимость уравнения регрессии** — значит *установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной.*

Проверка значимости уравнения регрессии производится на основе дисперсионного анализа. В гл. 11 дисперсионный анализ рассмотрен как самостоятельный инструмент (метод) статистического анализа. Здесь же он применяется как вспомогательное средство для изучения качества регрессионной модели.

Согласно основной идее дисперсионного анализа (см. гл. 11)

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_{x_i}-\bar{y})^2 + \sum_{i=1}^{n}(y_i-y_{x_i})^2 \qquad \textbf{(13.16)}$$

или $$Q = Q_R + Q_e, \qquad \textbf{(13.17)}$$

где Q — общая сумма квадратов отклонений зависимой переменной от средней, а Q_R и Q_e — соответственно сумма квадратов, обусловленная регрессией, и остаточная сумма квадратов, характеризующая влияние неучтенных факторов.

Убедимся в том, что пропущенное в **(13.17)** третье слагаемое $Q_3 = 2\sum_{i=1}^{n}(y_{x_i}-\bar{y})(y_i-y_{x_i})$ равно нулю. Учитывая **(13.7)** и первое уравнение системы **(12.11)**, имеем:

$$y_{x_i}-\bar{y} = b_1(x_i-\bar{x});$$

$$y_i - y_{x_i} = y_i - b_0 - b_1x_i = y_i - (\bar{y}-b_1\bar{x}) - b_1x_i = (y_i-\bar{y}) - b_1(x_i-\bar{x}).$$

Теперь

$$Q_3 = 2\sum_{i=1}^{n}(y_{x_i}-\bar{y})(y_i-y_{x_i}) = 2b_1\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) - 2b_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2 = 0$$

(с учетом соотношения **(13.10)**).

Схема дисперсионного анализа имеет вид, представленный в табл. 13.3.

Таблица 13.3

Компоненты дисперсии	*Сумма квадратов*	*Число степеней свободы*	*Средние квадраты*
Регрессия	$Q_R = \sum_{i=1}^{n}(y_{x_i}-\bar{y})^2$	$m-1$	$s^2_R = \frac{Q_R}{m-1}$
Остаточная	$Q_e = \sum_{i=1}^{n}(y_i-y_{x_i})^2$	$n-m$	$s^2 = \frac{Q_e}{n-m}$
Общая	$Q = \sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$	

Средние квадраты s^2_k и s^2 (табл. 13.3) представляют собой несмещенные оценки дисперсий зависимой переменной, обусловленной соответственно регрессией или объясняющей(ими) переменной(ыми) X и воздействием неучтенных случайных факторов и ошибок; m — число оцениваемых параметров уравнения регрессии; n — число наблюдений.

Замечание. При расчете общей суммы квадратов полезно иметь в виду, что

$$Q = \sum_{i=1}^{n}y_i^2 - \frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n} \qquad \textbf{(13.17′)}$$

(формула **(13.17′)** следует из разложения

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

учитывая, что $\bar{y}=\frac{\sum_{i=1}^{n}y_i}{n}$).

При отсутствии линейной зависимости между зависимой и объясняющей(ими) переменной(ыми) случайные величины $s_R^2=Q_R/(m-1)$ и $s^2=Q_e/(n-m)$ имеют χ^2-распределение соответственно с $m-1$ и $n-m$ степенями свободы, а их отношение — F-распределение с теми же степенями свободы (см. § 4.9). Поэтому уравнение регрессии значимо на уровне α, если фактически наблюдаемое значение статистики

$$F=\frac{Q_R(n-m)}{Q_e(m-1)}=\frac{s_R^2}{s^2}>F_{\alpha;k_1;k_2}, \qquad \textbf{(13.18)}$$

где $F_{\alpha;k_1;k_2}$ — табличное значение F-критерия Фишера—Снедекора, определенное на уровне значимости α при $k_1=m-1$ и $k_2=n-m$ степенях свободы.

Учитывая смысл величин s_R^2 и s^2, можно сказать, что *значение F показывает, в какой мере регрессия лучше оценивает значение зависимой переменной по сравнению с ее средней.*

В случае линейной парной регрессии $m=2$ и уравнение регрессии значимо на уровне α, если

$$F=\frac{Q_R(n-2)}{Q_e}>F_{\alpha;1;n-2}. \qquad \textbf{(13.18')}$$

Выше (§ 12.6) введен индекс корреляции R (для парной линейной модели — коэффициент корреляции r), выраженный через дисперсии (см. формулу **(12.60)**). Тот же коэффициент в терминах «сумм квадратов» примет вид:

$$R=\sqrt{\frac{Q_R}{Q}}=\sqrt{1-\frac{Q_e}{Q}}. \qquad \textbf{(13.19)}$$

Следует отметить, что значимость уравнения парной линейной регрессии может быть проверена и другим способом, если оценить **значимость коэффициента регрессии** b_1. Можно показать, что при выполнении предпосылки **5** регрессионного анализа (с. 458) статистика $t=\frac{b_1-\beta_1}{\sigma_{b_1}}$ имеет стандартный нормальный закон распределения $N(0;1)$, а если в выражении **(13.11)** для σ_{b_1} заменить параметр σ^2 его оценкой s^2, то статистика

$$t=\frac{b_1-\beta_1}{s}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad \textbf{(13.19')}$$

имеет t-распределение с $k=n-2$ степенями свободы. Поэтому коэффициент регрессии b_1 значим на уровне α, если $|t|=\frac{b_1}{s}\sqrt{\sum_{i=1}^{n}(x_i-x)^2}>t_{1-\alpha;n-2}$, а **доверительный интервал** для β_1 имеет вид:

$$b_1-t_{1-\alpha;n-2}\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\le\beta_1\le b_1+t_{1-\alpha;n-2}\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}. \qquad \textbf{(13.19'')}$$

Для парной регрессионной модели оценка значимости уравнения регрессии по F-критерию равносильна оценке значимости коэффициента регрессии b_1 либо коэффициента корреляции r по t-критерию (см. § 12.5), ибо эти критерии связаны соотношением $F=t^2$. А интервальные оценки для параметра β_1 **(13.19'')** — при нормальном законе распределения зависимой переменной и $\beta_{yx}=\beta_1$ **(12.51)** совпадают.

При построении **доверительного интервала для дисперсии возмущений** σ^2 исходят из того, что статистика $\frac{ns^2}{\sigma^2}$ имеет χ^2-распределение с $k=n-2$ степенями свободы. Поэтому интервальная оценка для σ^2 на уровне значимости α имеет вид (см. **(9.47)**):

$$\frac{ns^2}{\chi^2_{\alpha/2;n-2}}\le\sigma^2\le\frac{ns^2}{\chi^2_{1-\alpha/2;n-2}}. \qquad \textbf{(13.20)}$$

▷ **Пример 13.2.** По данным табл. 13.1 оценить на уровне $\alpha = 0{,}05$ значимость уравнения регрессии Y по X. Найти интервальную оценку для параметров β_1 и σ^2.

Р е ш е н и е. *1-й способ.* Выше, в примере **13.1**, были найдены $\sum_{i=1}^{10} y_i = 68$, $\sum_{i=1}^{10} y_i^2 = 496$.

Вычислим необходимые суммы квадратов по формулам **(13.16)**, **(13.17′)**:

$$Q = \sum_{i=1}^{10}(y_i - \bar{y})^2 = \sum_{i=1}^{10} y_i^2 - \frac{\left(\sum_{i=1}^{10} y_i\right)^2}{10} = 496 - \frac{68^2}{10} = 33{,}6;$$

$$Q_e = \sum_{i=1}^{10}(y_{x_i} - \bar{y})^2 = \sum_{i=1}^{10} e_i^2 = 8{,}39 \text{ (см. табл. 13.2);}$$

$$Q_R = Q - Q_e = 33{,}6 - 8{,}39 = 25{,}21.$$

По формуле **(13.18′)** $F = \dfrac{25{,}21(10-2)}{8{,}39} = 24{,}04$.

По таблице F-распределения (табл. VI приложений) $F_{0{,}05;1;8} = 4{,}20$. Так как $F > F_{0{,}05;1;8}$, то уравнение регрессии значимо.

2-й способ. Учитывая, что $b_1 = 1{,}016$, $\sum_{i=1}^{10}(x_i - \bar{x})^2 = 24{,}40$, $s^2 = 1{,}049$ (см. пример **13.1**, табл. 13.2), по формуле **(13.19′)** $t = \dfrac{1{,}016}{\sqrt{1{,}049}}\sqrt{24{,}40} = 4{,}90$.

По таблице t-распределения (табл. IV приложений) $t_{0{,}95;8} = 2{,}31$. Так как $t > t_{0{,}95;8}$, то коэффициент регрессии, а значит, и уравнение парной линейной регрессии Y по X значимы. Оба способа оценки значимости уравнения парной регрессии равносильны, ибо $F = t^2$ $(24{,}40 = 4{,}90^2)$.

Найдем $100(1 - \alpha) = 95\%$-ный доверительный интервал для параметра β_1. По формуле **(13.19″)**

$$1{,}016 - 2{,}31\sqrt{\frac{1{,}049}{24{,}40}} \le \beta_1 \le 1{,}016 + 2{,}31\sqrt{\frac{1{,}049}{24{,}40}}$$

или $0{,}537 \le \beta_1 \le 1{,}495$, т.е. с надежностью 0,95 при изменении мощности пласта X на 1 м суточная выработка Y будет изменяться на величину, заключенную в интервале от 0,537 до 1,495 (т).

Найдем 95%-ный интервал для параметра σ^2.

Учитывая, что $\alpha = 1 - 0{,}95 = 0{,}05$, найдем по табл. V приложений $\chi^2_{\alpha/2;n-2} = \chi^2_{0{,}025;8} = 17{,}53$; $\chi^2_{1-\alpha/2;n-2} = \chi^2_{0{,}975;8} = 2{,}18$. По формуле **(13.20)**

$$\frac{10 \cdot 1{,}049}{17{,}5} \le \sigma^2 \le \frac{10 \cdot 1{,}049}{2{,}18} \text{ или } 0{,}599 \le \sigma^2 \le 4{,}81 \text{ и } 0{,}774 \le \sigma \le 2{,}19.$$

Таким образом, с надежностью 0,95 дисперсия возмущений заключена в пределах от 0,599 до 4,81, а их стандартное отклонение — от 0,774 до 2,19 (т). ▶

13.4. Нелинейная регрессия

Соотношения между социально-экономическими явлениями и процессами далеко не всегда можно выразить линейными функциями, так как при этом могут возникать неоправданно большие ошибки. В таких случаях используют нелинейную (по объясняющей переменной) регрессию.

Выбор вида уравнения регрессии **(8.3)** (этот важный этап анализа называется *спецификацией* или *этапом параметризации модели*) производится на основании опыта предыдущих исследований, литературных источников, других соображений профессионально-теоретического характера, а также визуального наблюдения расположения точек корреляционного поля. Наиболее часто встречаются следующие виды уравнений нелинейной регрессии: *полиномиальное* $y_x = b_0 + b_1 x + \ldots + b_k x^k$, *гиперболическое* $y_x = b_0 + b_1/x$, *степенное* $y_x = b_0 \cdot x_1^{b_1} \cdot \ldots \cdot x_p^{b_p}$.

Например, если исследуемый экономический показатель y при росте объема производства x состоит из двух частей — постоянной (не зависящей от x) и переменной (уменьшающейся с ростом x), то зависимость y от x можно представить в виде гиперболы $y_x = b_0 + b_1/x$. Если же показатель y отражает экономический процесс, который под влиянием фактора x происходит с постоянным ускорением или замедлением, то применяются полиномы. В ряде случаев для описания экономических процессов используются более сложные функции. Например, если процесс вначале ускоренно развивается, а затем, после достижения некоторого уровня, затуха-

ет и приближается к некоторому пределу, то могут оказаться полезными логистические функции типа $y = b_0/(1 + b_1 b_2^{f(x)})$.

При исследовании степенного уравнения регрессии следует иметь в виду, что оно нелинейно относительно параметров b_j, однако путем логарифмирования может быть преобразовано в линейное:

$$\ln y_x = \ln b_0 + b_1 \ln x_1 + \ldots + b_p \ln x_p.$$

Для определения неизвестных параметров b_0, b_1, ..., b_p, как и ранее, используется метод наименьших квадратов.

▷ **Пример 13.3.** По данным табл. 13.4 исследовать зависимость урожайности зерновых культур Y (ц/га) от количества осадков X (см), выпавших в вегетационный период[1].

Таблица 13.4

№ п/п	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Количество осадков x_i (см)	25	27	30	35	36	38	39	41	42	45	46	47	50	52	53
Урожайность y_i (ц/га)	23	24	27	27	32	31	33	35	34	32	29	28	25	24	25

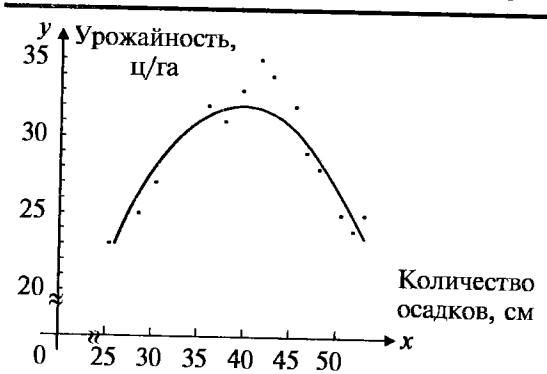


Рис. 13.3

Р е ш е н и е. Из качественных соображений можно предположить, что увеличение количества выпавших осадков приводит к увеличению урожайности до некоторого предела, после чего урожайность будет снижаться. Учитывая, кроме того, расположение точек корреляционного поля (см. рис. 13.3), можно предположить, что наиболее подходящим уравнением регрессии будет уравнение параболы

$$y_x = b_0 + b_1 x + b_2 x^2.$$

Его параметры b_0, b_1, b_2 находим, применяя метод наименьших квадратов:

$$S = \sum_{i=1}^{n} (y_{x_i} - \bar{y}_i)^2 = \sum_{i=1}^{n} (b_0 + b_1 x_i + b_2 x_i^2 - \bar{y}_i)^2 \to \min.$$

Приравняв частные производные $\frac{\partial s}{\partial b_0}$, $\frac{\partial s}{\partial b_1}$ и $\frac{\partial s}{\partial b_2}$ к нулю, получим после преобразований систему нормальных уравнений:

$$\begin{cases} b_0 n + b_1 \sum\limits_{i=1}^{n} x_i + b_2 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} y_i, \\ b_0 \sum\limits_{i=1}^{n} x_i + b_1 \sum\limits_{i=1}^{n} x_i^2 + b_2 \sum\limits_{i=1}^{n} x_i^3 = \sum\limits_{i=1}^{n} y_i x_i, \\ b_0 \sum\limits_{i=1}^{n} x_i^2 + b_1 \sum\limits_{i=1}^{n} x_i^3 + b_2 \sum\limits_{i=1}^{n} x_i^4 = \sum\limits_{i=1}^{n} y_i x_i^2. \end{cases} \quad (13.21)$$

Для расчета необходимых сумм составим вспомогательную таблицу (табл. 13.5).

Таблица 13.5

i	x_i	y_i	x_i^2	x_i^3	x_i^4	$y_i x_i$	$y_i x_i^2$	y_i^2	$\bar{y}_{x_i}$	$(y_{x_i} - y_i)^2$
1	25	23	625	15 625	390 625	575	14 375	529	21,7	1,69
2	27	24	729	19 683	531 441	648	17 496	576	24,3	0,11
...	...	...	...	...	...	...	...	...	...	...
14	52	24	2704	140 608	7 311 616	1248	64 896	576	24,7	0,46
15	53	25	2809	148 877	7 890 481	1325	70 225	625	23,4	2,44
$\sum$	606	429	25 548	1 115 808	50 158 200	17 371	730 123	12 493	—	45,94

[1] То есть в период роста, развития растений.

Теперь система **(13.21)** примет вид:

$$\begin{cases} 15b_0 + 606b_1 + 25\,548b_2 = 429, \\ 606b_0 + 25\,548b_1 + 1\,115\,808b_2 = 17\,371, \\ 25\,548b_0 + 1\,115\,808b_1 + 50\,158\,200b_2 = 730\,123. \end{cases}$$

Решая эту систему, например, методом Гаусса, получим $b_0=-43{,}93$; $b_1=3{,}8342$; $b_2=-0{,}048361$, т.е. уравнение регрессии имеет вид:

$$y_x=-43{,}93+3{,}8342x-0{,}048361x^2.$$

Оценим значимость полученной зависимости. С этой целью по **(13.16)** найдем суммы (см. итоговую строку табл. 13.5):

$$Q_e=\sum_{i=1}^{n}(y_{x_i}-y_i)^2=45{,}94;$$

$$Q=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}=12\,493-\frac{429^2}{15}=223{,}6;$$

$$Q_R=Q-Q_e=223{,}6-45{,}94=177{,}66.$$

По **(13.18)**: $F=\dfrac{177{,}66\cdot(15-3)}{45{,}94\cdot(3-1)}=23{,}2$. Табличное значение $F_{0{,}05;2;12}=3{,}88$. Так как $F>F_{0{,}05;2;12}$, то уравнение регрессии значимо.

Для оценки тесноты связи вычислим индекс корреляции по формуле **(13.19)**:

$$R_{yx}=\sqrt{1-\frac{Q_e}{Q}}=\sqrt{1-\frac{45{,}94}{223{,}6}}=\sqrt{0{,}795}=0{,}891,$$

т.е. полученная зависимость весьма тесная. Коэффициент детерминации $R^2_{yx}=0{,}795$ показывает, что вариация урожайности зерновых культур на 79,5% обусловлена регрессией, или изменчивостью количества выпавших в вегетационный период осадков. ▶

В некоторых случаях нелинейность связей является следствием качественной неоднородности совокупности, к которой применяют регрессионный анализ. Например, объединение в одной совокупности предприятий различной специализации или

предприятий, существенно различающихся по природным условиям, и т.д. В этих случаях нелинейность может являться следствием механического объединения разнородных единиц. Регрессионный анализ таких совокупностей не может быть эффективным. Поэтому любая нелинейность связей должна критически анализироваться.

По расположению точек корреляционного поля далеко не всегда можно принять окончательное решение о виде уравнения регрессии. Если теоретические соображения или опыт предыдущих исследований не могут подсказать точного решения, то необходимо сделать расчеты по двум или нескольким уравнениям. Предпочтение отдается уравнению, для которого меньше величина *остаточной дисперсии*. Однако при незначительных расхождениях в остаточных дисперсиях следует всегда останавливаться на более простом уравнении, интерпретация показателей которого не представляется сложной.

Весьма заманчивым представляется увеличение порядка выравнивающей параболической кривой, ибо известно, что всякую функцию на любом интервале можно как угодно точно приблизить полиномом $y=b_0+b_1x+b_2x^2+...+b_kx^k$. Так, можно подобрать такой показатель k, что соответствующий полином пройдет через все вершины эмпирической линии регрессии. Однако повышение порядка гипотетической параболической кривой может привести к неоправданному усложнению вида искомой функции регрессии, когда случайные отклонения осредненных точек неправильно истолковываются как определенные закономерности в поведении кривой регрессии. Кроме того, за счет увеличения числа параметров снижается точность кривой регрессии (особенно в случае малой по объему выборки) и увеличивается объем вычислительных работ. В связи с этим в практике регрессионного анализа для выравнивания крайне редко используются полиномы выше третьей степени.

13.5. Множественный регрессионный анализ

Экономические явления, как правило, определяются большим числом одновременно и совокупно действующих факторов. В связи с этим часто возникает задача исследования зависимости одной зависимой переменной Y от нескольких объясняющих переменных $X_1,X_2,...,X_n$. Эта задача решается с помощью *множественного регрессионного анализа.*

Обозначим i-е наблюдение переменной y_i, а объясняющих переменных — $x_{i1},x_{i2},...,x_{ip}$. Тогда *модель множественной линейной регрессии* можно представить в виде:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i, \qquad \textbf{(13.22)}$$

где $i=1,2,...,n$, а ε_i удовлетворяет приведенным выше предпосылкам **(13.3)—(13.5)**.

Включение в регрессионную модель новых объясняющих переменных усложняет получаемые формулы и вычисления. Это приводит к целесообразности использования матричных обозначений. Матричное описание регрессии облегчает как теоретические концепции анализа, так и необходимые расчетные процедуры.

Введем обозначения: $Y = (y_1\ y_2\ ...\ y_n)'$ — *матрица-столбец, или вектор, значений зависимой переменной размера* n[1];

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ ..... & & & & \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix}$$

— *матрица значений объясняющих переменных*, или *матрица плана* размера $n\times(p+1)$ (обращаем внимание на то, что в матрицу X дополнительно введен столбец, все элементы которого равны 1, т.е. условно полагается, что в модели **(13.22)** свободный член β_0 умножается на фиктивную переменную x_{i0}, принимающую значение 1 для всех i: $x_{i0} \equiv 1$ $(i=1,2,...,n)$;

$\beta = (\beta_0\ \beta_1\ ...\ \beta_p)'$ — *матрица-столбец, или вектор, параметров* размера $(p+1)$;

$\varepsilon = (\varepsilon_1\ \varepsilon_2\ ...\ \varepsilon_n)'$ — *матрица-столбец, или вектор, возмущений (случайных ошибок, остатков) размера* n.

Тогда в матричной форме модель **(13.22)** примет вид:

$$Y = X\beta + \varepsilon. \qquad \textbf{(13.23)}$$

[1] Знаком ' обозначается операция транспонирования матриц.

Оценкой этой модели по выборке является уравнение

$$Y = Xb + e,$$

где $b = (b_0\ b_1\ ...\ b_p)'$, $e = (e_1\ e_2\ ...\ e_n)'$.

Для оценки вектора неизвестных параметров β применим **метод наименьших квадратов**. Так как произведение транспонированной матрицы e' на саму матрицу e

$$e'e = (e_1\ e_2\ ...\ e_n)\begin{pmatrix} e_1 \\ e_2 \\ ... \\ e_n \end{pmatrix} = e_1^2 + e_2^2 + ... + e_n^2 = \sum_{i=1}^{n} e_i^2,$$

то условие минимизации остаточной суммы квадратов запишется в виде:

$$S = \sum_{i=1}^{n} (y_{x_i} - y_i)^2 = \sum_{i=1}^{n} e_i^2 = e'e = (Y - Xb)'(Y - Xb) \to \min. \qquad \textbf{(13.24)}$$

Учитывая, что при транспонировании произведения матриц получается произведение транспонированных матриц, взятых в обратном порядке, т.е. $(Xb)'=b'X'$, получим после раскрытия скобок:

$$S = Y'Y - b'X'Y - Y'Xb + b'X'Xb.$$

Произведение $Y'Xb$ есть матрица размера $(1\times n)[n\times(p+1)\times\times[(p+1)\times 1]=(1\times 1)$, т.е. величина скалярная, следовательно, оно не меняется при транспонировании: $Y'Xb=(Y'Xb)'=b'X'Y$. Поэтому условие минимизации **(13.24)** примет вид:

$$S = Y'Y - 2b'X'Y + b'X'Xb \to \min. \qquad \textbf{(13.24')}$$

На основании необходимого условия экстремума функции нескольких переменных $S(b_0,b_1,...,b_p)$, представляющей **(13.24)**, необходимо приравнять к нулю частные производные по этим переменным или в матричной форме — вектор частных производных

$$\frac{\partial S}{\partial b} = \left(\frac{\partial S}{\partial b_0}\ \frac{\partial S}{\partial b_1}\ ...\ \frac{\partial S}{\partial b_p} \right).$$

Для вектора частных производных доказаны следующие формулы[1]:

$$\frac{\partial}{\partial b}(b'c)=c,\quad \frac{\partial}{\partial b}(b'Ab)=2Ab,$$

где b и c — вектор-столбцы, а A — симметрическая матрица, в которой элементы, расположенные симметрично относительно главной диагонали, равны.

Поэтому, полагая $c = X'Y$, а матрицу $A = X'X$ (она является симметрической — см. **(13.26)**), найдем

$$\frac{\partial S}{\partial b}=-2X'Y'+2X'Xb=0,$$

откуда получаем систему нормальных уравнений в матричной форме для определения вектора b:

$$X'Xb=X'Y. \qquad \textbf{(13.25)}$$

Найдем матрицы, входящие в это уравнение[2]. Матрица $X'X$ представляет матрицу сумм первых степеней, квадратов и попарных произведений n наблюдений объясняющих переменных:

$$X'X=\begin{pmatrix}1 & 1 & \dots & 1\\ x_{11} & x_{21} & \dots & x_{n1}\\ \dots & \dots & \dots & \dots\\ x_{1p} & x_{2p} & \dots & x_{np}\end{pmatrix}\begin{pmatrix}1 & x_{11} & \dots & x_{1p}\\ 1 & x_{21} & \dots & x_{2p}\\ \dots & \dots & & \dots\\ 1 & x_{n1} & \dots & x_{np}\end{pmatrix}= \qquad \textbf{(13.26)}$$

$$=\begin{pmatrix}n & \sum x_{i1} & \dots & \sum x_{ip}\\ \sum x_{i1} & \sum x_{i1}^2 & \dots & \sum x_{i1}x_{ip}\\ \dots & \dots & & \dots\\ \sum x_{ip} & \sum x_{i1}x_{ip} & \dots & \sum x_{ip}^2\end{pmatrix}.$$

[1] Справедливость приведенных формул проиллюстрируем на примере.

Пусть $b=\begin{pmatrix}b_1\\ b_2\end{pmatrix}$, $c=\begin{pmatrix}3\\ 4\end{pmatrix}$, $A=\begin{pmatrix}2 & 3\\ 3 & 5\end{pmatrix}$. Так как $b'c=(b_1\ b_2)\begin{pmatrix}3\\ 4\end{pmatrix}=3b_1+4b_2$ и

$b'Ab=(b_1b_2)\begin{pmatrix}2 & 3\\ 3 & 5\end{pmatrix}\begin{pmatrix}b_1\\ b_2\end{pmatrix}=2b_1^2+6b_1b_2+5b_2^2$, то $\frac{\partial}{\partial b}(b'c)=\frac{\partial}{\partial b}(3b_1+4b_2)=\begin{pmatrix}3\\ 4\end{pmatrix}=c$

и $\frac{\partial}{\partial b}(b'Ab)=\frac{\partial}{\partial b}(2b_1^2+6b_1b_2+5b_2^2)=\begin{pmatrix}4b_1+6b_2\\ 6b_1+10b_2\end{pmatrix}=2\begin{pmatrix}2 & 3\\ 3 & 5\end{pmatrix}\begin{pmatrix}b_1\\ b_2\end{pmatrix}=2Ab$

[2] Здесь под знаком $\sum$ подразумевается $\sum_{i=1}^{n}$.

Матрица $X'Y$ есть вектор произведений n наблюдений объясняющих и зависимой переменных:

$$X'Y=\begin{pmatrix}1 & 1 & \dots & 1\\ x_{11} & x_{21} & \dots & x_{n1}\\ \dots & \dots & \dots & \dots\\ x_{1p} & x_{2p} & \dots & x_{np}\end{pmatrix}\begin{pmatrix}y_1\\ y_2\\ \dots\\ y_n\end{pmatrix}=\begin{pmatrix}\sum y_i\\ \sum y_ix_{i1}\\ \dots\\ \sum y_ix_{ip}\end{pmatrix}. \qquad \textbf{(13.27)}$$

В частном случае из рассматриваемого матричного уравнения **(13.25)** с учетом **(13.26)** и **(13.27)** для одной объясняющей переменной ($p = 1$) нетрудно получить уже рассматриваемую систему нормальных уравнений **(12.10)** для несгруппированных данных. Действительно, в этом случае матричное уравнение **(13.25)** принимает вид[1]:

$$\begin{pmatrix}n & \sum x_i\\ \sum x_i & \sum x_i^2\end{pmatrix}\begin{pmatrix}b_0\\ b_1\end{pmatrix}=\begin{pmatrix}\sum y_i\\ \sum y_ix_i\end{pmatrix},$$

откуда непосредственно следует система нормальных уравнений **(12.10)** для несгруппированных данных.

Для решения матричного уравнения **(13.25)** относительно вектора оценок параметров b необходимо ввести еще одну **предпосылку 6** для множественного регрессионного анализа: *матрица $X'X$ является неособенной*, т.е. ее определитель не равен нулю. Следовательно, ранг матрицы $X'X$ равен ее порядку, т.е. $r(X'X)=p+1$. Из матричной алгебры известно (см., например, [9]), что $r(X'X)=r(X)$, значит, $r(X)=p+1$, т.е. ранг матрицы плана X равен числу ее столбцов. Это позволяет сформулировать предпосылку **6** множественного регрессионного анализа в следующем виде:

6. *Векторы значений объясняющих переменных, или столбцы матрицы плана X, должны быть линейно независимыми, т.е. ранг матрицы X — максимальный* ($r(X) = p + 1$).

Кроме того, полагают, что число имеющихся наблюдений (значений) каждой из объясняющих переменных превосходит ранг матрицы X, т.е. $n > r(X)$ или $n > p + 1$, ибо в противном

[1] В случае одной объясняющей переменной отпадает необходимость в записи под символом x второго индекса, указывающего номер переменной.

случае в принципе невозможно получение сколько-нибудь надежных статистических выводов.

Решением уравнения **(13.25)** является вектор

$$b=(X'X)^{-1}X'Y, \qquad \textbf{(13.28)}$$

где $(X'X)^{-1}$ — *матрица, обратная матрице коэффициентов системы* **(13.25)**, а $X'Y$ — *матрица-столбец*, или *вектор, ее свободных членов.*

Зная вектор b, выборочное уравнение множественной регрессии представим в виде:

$$y_{x_0}=X_0'b, \qquad \textbf{(13.29)}$$

где y_{x_0} — групповая (условная) средняя переменной Y при заданном векторе значений объясняющей переменной $X_0'=(1 \quad x_{10} \quad x_{20} \quad \ldots \quad x_{p0})$.

▷ **Пример 13.4.** Имеются следующие данные[1] (условные) о сменной добыче угля на одного рабочего Y (т), мощности пласта X_1 (м) и уровне механизации работ X_2 (%), характеризующие процесс добычи угля в 10 шахтах.

Таблица 13.6

i	x_{i1}	x_{i2}	y_i	i	x_{i1}	x_{i2}	y_i
1	8	5	5	6	8	8	6
2	11	8	10	7	9	6	6
3	12	8	10	8	9	4	5
4	9	5	7	9	8	5	6
5	8	7	5	10	12	7	8

Предполагая, что между переменными Y, X_1 и X_2 существует линейная корреляционная зависимость, найти ее аналитическое выражение (уравнение регрессии Y по X_1 и X_2).

[1] В этом примере использованы данные примера **13.1** с добавлением результатов наблюдений над новой объясняющей переменной X_2, при этом старую переменную X из примера **13.1** обозначаем теперь X_1.

Р е ш е н и е. Обозначим

$$Y=\begin{pmatrix}5\\10\\\ldots\\8\end{pmatrix}, \quad X=\begin{pmatrix}1 & 8 & 5\\1 & 11 & 8\\\ldots & \ldots & \ldots\\1 & 12 & 7\end{pmatrix}$$

(напоминаем, что в матрицу плана X вводится дополнительный столбец чисел, состоящий из единиц).

Для удобства вычислений составляем вспомогательную таблицу.

Таблица 13.7

i	x_{i1}	x_{i2}	y_i	x_{i1}^2	x_{i2}^2	y_i^2	$x_{i1}x_{i2}$	y_ix_{i1}	y_ix_{i2}	y_{x_i}	$e_i^2=(y_{x_i}-y_i)^2$
1	8	5	5	64	25	25	40	40	25	5,13	0,016
2	11	8	10	121	64	100	88	110	80	8,79	1,464
3	12	8	10	144	64	100	96	120	80	9,64	1,127
4	9	5	7	81	25	49	45	63	35	5,98	1,038
5	8	7	5	64	49	25	56	40	35	5,86	0,741
6	8	8	6	64	64	36	64	48	48	6,23	0,052
7	9	6	6	81	36	36	54	54	36	6,35	0,121
8	9	4	5	81	16	25	36	45	20	5,61	0,377
9	8	5	6	64	25	36	40	48	30	5,13	0,762
10	12	7	8	144	49	64	84	96	56	9,28	1,631
Σ	94	63	68	908	417	496	603	664	445	—	6,329

Теперь

$$X'X=\begin{pmatrix}1 & 1 & \ldots & 1\\8 & 11 & \ldots & 12\\5 & 8 & \ldots & 7\end{pmatrix}\begin{pmatrix}1 & 8 & 5\\1 & 11 & 8\\\ldots & \ldots & \ldots\\1 & 12 & 7\end{pmatrix}=\begin{pmatrix}10 & 94 & 63\\94 & 908 & 603\\63 & 603 & 417\end{pmatrix}$$

(см. суммы в итоговой строке табл. 13.7);

$$X'Y=\begin{pmatrix}1 & 1 & \ldots & 1\\8 & 11 & \ldots & 12\\5 & 8 & \ldots & 7\end{pmatrix}\begin{pmatrix}5\\10\\\ldots\\8\end{pmatrix}=\begin{pmatrix}68\\664\\445\end{pmatrix}.$$

Матрицу $A^{-1}=(X'X)^{-1}$ определим по формуле $A^{-1}=\frac{1}{|A|}\overline{A}$, где $|A|$ — определитель матрицы $X'X$, $\overline{A}$ — матрица, присоединенная к матрице $X'X$. Получим (рекомендуем читателю убедиться в этом самостоятельно)

$$A^{-1}=\frac{1}{3738}\begin{pmatrix}15\,027 & -1209 & -522\\ -1209 & 201 & -108\\ -522 & -108 & 244\end{pmatrix}.$$

Теперь в соответствии с **(13.28)**, умножая эту матрицу на вектор

$$X'Y=\begin{pmatrix}68\\664\\445\end{pmatrix},\ \text{получим}\ b=\frac{1}{3738}\begin{pmatrix}-13\,230\\3192\\1372\end{pmatrix}=\begin{pmatrix}-3{,}5393\\0{,}8539\\0{,}3670\end{pmatrix}.$$

С учетом **(13.29)** уравнение множественной регрессии имеет вид: $y_x=-3{,}54+0{,}854x_1+0{,}367x_2$. Оно показывает, что при увеличении только мощности пласта X_1 (при неизменном X_2) на 1 м, добыча угля на одного рабочего Y увеличивается в среднем на 0,854 т, а при увеличении только уровня механизации работ X_2 (при неизменной X_1) — в среднем на 0,367 т.

Добавление в регрессионную модель новой объясняющей переменной X_2 изменило коэффициент регрессии b_1 (Y по X_1) с 1,016 для парной регрессии (см. пример **13.1**) до 0,854 — для множественной регрессии. В этом никакого противоречия нет, так как во втором случае коэффициент регрессии позволяет оценить прирост зависимой переменной Y при изменении на единицу объясняющей переменной X_1 в чистом виде, независимо от X_2. В случае парной регрессии b_1 учитывает воздействие на Y не только переменной X_1, но и косвенно корреляционно связанной с ней переменной X_2. ▶

На практике часто бывает необходимо сравнение влияния на зависимую переменную различных объясняющих переменных, когда последние выражаются разными единицами измерения. В этом случае используют *стандартизованные коэффициенты регрессии* b'_j и *коэффициенты эластичности* E_j (j=1,2,...,p):

$$b'_j=b_j\frac{s_{x_j}}{s_y};\qquad \textbf{(13.30)}$$

$$E_j=b_j\frac{\overline{x}_j}{\overline{y}}.\qquad \textbf{(13.31)}$$

Стандартизованный коэффициент регрессии b'_j показывает, на сколько величин s_y изменится в среднем зависимая переменная Y при увеличении только j-й объясняющей переменной на s_{x_j}, а коэффициент эластичности E_j — на сколько процентов (от средней) изменится в среднем Y при увеличении только X_j на 1%.

▷ **Пример 13.5.** По данным примера **13.4** сравнить раздельное влияние на сменную добычу угля двух факторов — мощности пласта и уровня механизации работ.

Р е ш е н и е. Для сравнения влияния каждой из объясняющих переменных по формуле **(13.30)** вычислим стандартизованные коэффициенты регрессии:

$$b'_1=0{,}8539\cdot\frac{1{,}56}{1{,}83}=0{,}728;\quad b'_2=0{,}3670\cdot\frac{1{,}42}{1{,}83}=0{,}285,$$

а по формуле **(13.31)** — коэффициенты эластичности:

$$E_1=0{,}8539\cdot\frac{9{,}4}{6{,}8}=1{,}180;\quad E_2=0{,}3670\cdot\frac{6{,}3}{6{,}8}=0{,}340.$$

(Здесь мы опустили расчет необходимых характеристик переменных:

$$\overline{x}_1=9{,}4;\ \overline{x}_2=6{,}3;\ \overline{y}=6{,}8;\ s_{x_1}=1{,}56;\ s_{x_2}=1{,}42;\ s_y=1{,}83\ .)$$

Таким образом, увеличение мощности пласта и уровня механизации работ только на одно s_{x_1} или на одно s_{x_2} увеличивает в среднем сменную добычу угля на одного рабочего соответственно на $0{,}728s_y$ или на $0{,}285s_y$, а увеличение этих переменных на 1% (от своих средних значений) приводит в среднем к росту добычи угля соответственно на 1,18% и 0,34%. Итак, по обоим показателям на сменную добычу угля большее влияние оказывает фактор «мощность пласта» по сравнению с фактором «уровень механизации работ». ▶

Преобразуем **вектор оценок (13.28)** с учетом **(13.23)**:

$$b=(X'X)^{-1}X'(X\beta+\varepsilon)=(X'X)^{-1}(X'X)\beta+(X'X)^{-1}X'\varepsilon=$$
$$=E\beta+(X'X)^{-1}X'\varepsilon$$

или

$$b=\beta+(X'X)^{-1}X'\varepsilon, \qquad \textbf{(13.32)}$$

т.е. *оценки параметров* **(13.28)**, найденные по выборке, *будут содержать случайные ошибки.*

13.6. Ковариационная матрица и ее выборочная оценка

Вариации оценок параметров будут в конечном счете определять точность уравнения множественной регрессии. Для их измерения в многомерном регрессионном анализе рассматривают так называемую *ковариационную матрицу K*, являющуюся матричным аналогом дисперсии одной переменной:

$$\overset{\bullet}{K}=\begin{pmatrix} & & & \\ K_{00} & K_{01} & \ldots & K_{0p} \\ K_{10} & K_{11} & \ldots & K_{1p} \\ \ldots & \ldots & \ldots & \ldots \\ K_{p0} & K_{p1} & \ldots & K_{pp} \end{pmatrix},$$

где элементы K_{ij} — *ковариации* (или *корреляционные моменты*) оценок параметров β_i и β_j ($i, j = 0, 1, \ldots, p$). Ковариация двух переменных определяется как математическое ожидание произведения отклонений этих переменных от их математических ожиданий (см. § 5.6). Поэтому

$$K_{ij}=M[(b_i-M(b_i))(b_j-M(b_j))]. \qquad \textbf{(13.33)}$$

Ковариация характеризует как степень рассеяния значений двух переменных относительно их математических ожиданий, так и взаимосвязь этих переменных.

В силу того, что оценки b_j, полученные методом наименьших квадратов, являются несмещенными оценками параметров β_j, т.е. $M(b_j)=\beta_j$, выражение **(13.33)** примет вид:

$$K_{ij}=M[(b_i-\beta_i)(b_j-\beta_j)].$$

Рассматривая ковариационную матрицу K, легко заметить, что на ее главной диагонали находятся дисперсии оценок параметров регрессии, ибо

$$K_{jj}=M[(b_j-\beta_j)\,(b_j-\beta_j)]=M(b_j-\beta_j)^2=\sigma^2_{b_j}. \qquad \textbf{(13.34)}$$

В сокращенном виде ковариационная матрица K имеет вид:

$$K=M[(b-\beta)\,(b-\beta)']$$

(в этом легко убедиться, перемножив векторы $(b-\beta)$ и $(b-\beta)'$

Учитывая **(13.32)**, преобразуем это выражение:

$$K=M\{[(X'X)^{-1}X'\varepsilon]\,[(X'X)^{-1}X'\varepsilon]'\}= \qquad \textbf{(13.35)}$$
$$=M[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}]=(X'X)^{-1}X'M(\varepsilon\varepsilon')X(X'X)^{-1},$$

ибо элементы матрицы X —неслучайные величины.

Матрица $M(\varepsilon\varepsilon')$ представляет собой ковариационную *матрицу вектора возмущений* ε:

$$M(\varepsilon\varepsilon')=\begin{pmatrix} M(\varepsilon_1^2) & M(\varepsilon_1\varepsilon_2) & \ldots & M(\varepsilon_1\varepsilon_n) \\ M(\varepsilon_2\varepsilon_1) & M(\varepsilon_2^2) & \ldots & M(\varepsilon_2\varepsilon_n) \\ \ldots & \ldots & \ldots & \ldots\ldots \\ M(\varepsilon_n\varepsilon_1) & M(\varepsilon_n\varepsilon_2) & \ldots & M(\varepsilon_n^2) \end{pmatrix},$$

в которой все элементы, не лежащие на главной диагонали, равны нулю в силу предпосылки **4** о некоррелированности возмущений ε_i и ε_j между собой (см. **(13.5)**), а все элементы, лежащие на главной диагонали, в силу предпосылок **2** и **3** регрессионного анализа (см. **(13.3)** и **(13.4)**) равны одной и той же дисперсии σ^2:

$$M(\varepsilon_i^2)=M(\varepsilon_i-0)^2=D(\varepsilon_i^2)=\sigma^2.$$

Поэтому матрица $M(\varepsilon\varepsilon')=\sigma^2E$, где E — единичная матрица n-го порядка. Следовательно, в силу **(13.35)** ковариационная матрица вектора b оценок параметров:

$$K=[(X'X)^{-1}X'(\sigma^2E)]X(X'X)^{-1}=\sigma^2(X'X)^{-1}(X'EX)(X'X)^{-1}$$

или

$$K=\sigma^2(X'X)^{-1} \qquad \textbf{(13.36)}$$

Итак, *с помощью обратной матрицы* $(X'X)^{-1}$ *определяется не только сам вектор b оценок параметров* **(13.28)**, *но и дисперсии и ковариации его компонент.*

Входящая в **(13.36)** дисперсия возмущений неизвестна. Заменив ее выборочной остаточной дисперсией

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)} = \frac{e'e}{n-p-1}, \qquad \textbf{(13.37)}$$

по **(13.36)** получаем выборочную оценку ковариационной матрицы K. (В знаменателе выражения **(13.37)** стоит $n-(p+1)$, а не $n-2$, как это было выше в **(13.6)**. Это связано с тем, что теперь $p+1$ степеней свободы (а не две) теряются при определении неизвестных параметров, число которых вместе со свободным членом b_0 равно $p+1$).

13.7. Определение доверительных интервалов для коэффициентов и функции регрессии

Перейдем теперь к оценке значимости коэффициентов регрессии b_j и построению доверительного интервала для параметров регрессионной модели β_j ($j=1,2,...,p$).

В силу **(13.34)**, **(13.36)** и изложенного выше оценка дисперсии коэффициента регрессии b_j определится по формуле:

$$s_{bj}^2 = s^2[(X'X)^{-1}]_{jj},$$

где s^2 — несмещенная оценка параметра σ^2;

$[(X'X)^{-1}]_{jj}$ — диагональный элемент матрицы $(X'X)^{-1}$.

Среднее квадратическое отклонение (стандартная ошибка) коэффициента регрессии b_j примет вид:

$$s_{b_j} = s\sqrt{[(X'X)^{-1}]_{jj}}. \qquad \textbf{(13.38)}$$

Значимость коэффициента регрессии b_j можно проверить, если учесть, что статистика $(b_j-\beta_j)/s_{b_j}$ имеет t-распределение Стьюдента с $k=n-p-1$ степенями свободы. Поэтому b_j значимо отличается от нуля на уровне значимости α, если

$|t| = (|b_j|/s_{b_j}) > t_{1-\alpha;k}$, а соответствующий $\gamma=(1-\alpha)\%$-ный доверительный интервал для параметра β_j есть

$$b_j - t_{1-\alpha;k}s_{b_j} \le \beta_j \le b_j + t_{1-\alpha;k}s_{b_j}. \qquad \textbf{(13.39)}$$

Наряду с интервальным оцениванием коэффициентов регрессии по **(13.39)** весьма важным для оценки точности определения зависимой переменной (прогноза) является построение **доверительного интервала для функции регрессии** или для условного математического ожидания зависимой переменной $M_{x_0}(Y)$, найденного в предположении, что объясняющие переменные $X_1, X_2, ..., X_p$ приняли значения, задаваемые вектором $X_0' = (1\ x_{10}\ x_{20}\ ...\ x_{p0})$. Выше такой интервал получен для уравнения парной регрессии (см. **(13.13)** и **(13.12)**). Обобщая соответствующие выражения на случай множественной регрессии, можно получить доверительный интервал для $M_{x_0}(Y)$:

$$y_{x_0} - t_{1-\alpha;k}s_{y_{x_0}} \le M_{x_0}(Y) \le y_{x_0} + t_{1-\alpha;k}s_{y_{x_0}}, \qquad \textbf{(13.40)}$$

где y_{x_0} — групповая средняя, определяемая по уравнению регрессии,

$$s_{y_{x_0}} = s\sqrt{X_0'(X'X)^{-1}X_0} \qquad \textbf{(13.41)}$$

— ее стандартная ошибка.

При обобщении формул **(13.15)** и **(13.14)** аналогичный **доверительный интервал для индивидуальных значений зависимой переменной** y_0^* примет вид:

$$y_{x_0} - t_{1-\alpha;n-p-1}s_{y_0} \le y_0^* \le y_{x_0} + t_{1-\alpha;n-p-1}s_{y_0}; \qquad \textbf{(13.42)}$$

где

$$s_{y_0} = s\sqrt{1 + X_0'(X'X)^{-1}X_0} \qquad \textbf{(13.43)}$$

Доверительный интервал для дисперсии возмущений σ^2 в множественной регрессии с надежностью $\gamma = 1-\alpha$ строится аналогично парной модели по формуле **(13.20)** с соответствующим изменением числа степеней свободы критерия χ^2:

$$\frac{ns^2}{\chi^2_{\alpha/2;n-p-1}} \le \sigma^2 \le \frac{ns^2}{\chi^2_{1-\alpha/2;n-p-1}} \qquad \textbf{(13.43')}$$

▷ **Пример 13.6.** По данным примера **13.4** оценить сменную добычу угля на одного рабочего для шахт с мощностью пласта 8 м и уровнем механизации работ 6%; найти 95%-ные доверительные интервалы для индивидуального и среднего значений сменной добычи угля на 1 рабочего для таких же шахт. Проверить значимость коэффициентов регрессии и построить для них 95%-ные доверительные интервалы. Найти с надежностью 0,95 интервальную оценку для дисперсии возмущений σ^2.

Р е ш е н и е. В примере **13.4** уравнение регрессии получено в виде: $y_x = -3{,}54 + 0{,}854x_1 + 0{,}367x_2$. По условию надо оценить $M_{x_0}(Y)$, где $X_0' = (1\ 8\ 6)$. Выборочной оценкой $M_{x_0}(Y)$ является групповая средняя, которую найдем по уравнению регрессии: $y_{x_0} = -3{,}54 + 0{,}854 \cdot 8 + 0{,}367 \cdot 6 = 5{,}49$ (т). Для построения доверительного интервала для $M_{x_0}(Y)$ необходимо знать дисперсию его оценки — $s^2_{y_{x_0}}$. Для ее вычисления обратимся к табл. 13.7 (точнее к ее двум последним столбцам, при составлении которых учтено, что групповые средние определяются по полученному уравнению регрессии).

Теперь по **(13.37)**: $s^2 = \dfrac{6{,}329}{10-2-1} = 0{,}904$ и $s = \sqrt{0{,}904} = 0{,}951$ (т).

Определяем стандартную ошибку групповой средней y_{x_0} по формуле **(13.41)**. Вначале найдем

$$X_0'(X'X)^{-1}X_0 = (1 \quad 8 \quad 6)\frac{1}{3738}\begin{pmatrix} 15\,027 & -1209 & -522 \\ -1209 & 201 & -108 \\ -522 & -108 & 244 \end{pmatrix}\begin{pmatrix} 1 \\ 8 \\ 6 \end{pmatrix} =$$

$$= \frac{1}{3738}(2223 \quad -249 \quad 78)\begin{pmatrix} 1 \\ 8 \\ 6 \end{pmatrix} = \frac{1}{3738}(699) = 0{,}1870.$$

Теперь $s_{y_{x_0}} = 0{,}951\sqrt{0{.}1870} = 0{,}411$ (т).

По табл. IV приложений при числе степеней свободы k=10−−2−1=7 находим $t_{0,95;7}$=2,36. По **(13.40)** доверительный интервал для $M_{x_0}(Y)$ равен $5{,}49 - 2{,}36 \cdot 0{,}411 \le M_{x_0}(Y) \le 5{,}49 + 2{,}36 \cdot 0{,}411$ или $4{,}52 \le M_{x_0}(Y) \le 6{,}46$ (т).

Итак, с надежностью 0,95 средняя сменная добыча угля на одного рабочего для шахт с мощностью пласта 8 м и уровнем механизации работ 6% находится в пределах от 4,52 до 6,46 т.

Сравнивая новый доверительный интервал для функции регрессии $M_{x_0}(Y)$, полученный с учетом двух объясняющих переменных, с аналогичным интервалом с учетом одной объясняющей переменной (см. пример **13.1**), можно заметить уменьшение его величины. Это связано с тем, что включение в модель новой объясняющей переменной позволяет несколько повысить точность модели за счет увеличения взаимосвязи зависимой и объясняющей переменных (см. ниже).

Найдем доверительный интервал для индивидуального значения y_0^* при $X_0' = (1\ 8\ 6)$:

по **(13.43)**: $s_{y_0} = 0{,}951\sqrt{1 + 0{,}1870} = 1{,}036$ (т) и по **(13.42)**: $5{,}49 - 2{,}36 \cdot 1{,}036 \le y_0^* \le 5{,}49 + 2{,}36 \cdot 1{,}036$, т.е. $3{,}05 \le y_0^* \le 7{,}93$ (т).

Итак, с надежностью 0,95 индивидуальное значение сменной добычи угля в шахтах с мощностью пласта 8 м и уровнем механизации работ 6% находится в пределах от 3,05 до 7,93 (т).

Проверим значимость коэффициентов регрессии b_1 и b_2. В примере **13.4** получены b_1=0,854 и b_2=0,367. Стандартная ошибка s_{b_1} в соответствии с **(13.38)** равна: $s_{b_1} = 0{,}951\sqrt{\dfrac{1}{3738} \cdot 201} = 0{,}221$.

Так как $t = \dfrac{0{,}854}{0{,}221} = 3{,}81 > t_{0,95;7} = 2{,}36$, то коэффициент b_1 значим.

Аналогично вычисляем $s_{b_2} = 0{,}951\sqrt{\dfrac{1}{3738} \cdot 244} = 0{,}243$ и

$t = \dfrac{0{,}367}{0{,}243} = 1{,}51 < t_{0,95;7} = 2{,}36$, т.е. коэффициент b_2 незначим на 5%-ном уровне.

Доверительный интервал имеет смысл построить только для значимого коэффициента регрессии b_1: по **(13.39)** $0{,}854 - -2{,}36 \cdot 0{,}221 \le \beta_1 \le 0{,}854 + 2{,}36 \cdot 0{,}221$ или $0{,}332 \le \beta_1 \le 1{,}376$.

Итак, с надежностью 0,95 за счет изменения на 1 м мощности пласта X_1 (при неизменном X_2) сменная добыча угля на одного рабочего Y будет изменяться в пределах от 0,332 до 1,376 т.

Найдем 95%-ный доверительный интервал для параметра σ^2. Учитывая, что $\alpha = 1 - 0{,}95 = 0{,}05$, $\alpha/2 = 0{,}025$, $1 - \alpha/2 = 0{,}975$, найдем по табл. V приложений при $n - p - 1 = n - 3$ степенях свободы $\chi^2_{0{,}025;7} = 16{,}0$; $\chi^2_{0{,}975;7} = 1{,}69$ и по формуле **(13.43′)**

$$\frac{10 \cdot 0{,}904}{16{,}0} \le \sigma^2 \le \frac{10 \cdot 0{,}904}{1{,}69} \text{ или } 0{,}565 \le \sigma^2 \le 5{,}35$$

$$\text{и } 0{,}751 \le \sigma^2 \le 2{,}31.$$

Таким образом, с надежностью 0,95 дисперсия возмущений заключена в пределах от 0,565 до 5,35, а их стандартное отклонение — от 0,751 до 2,31 (т). ▶

Формально переменные, имеющие незначимые коэффициенты регрессии, могут быть исключены из рассмотрения. В экономических исследованиях исключению переменных из регрессии должен предшествовать тщательный *качественный* анализ. Поэтому может оказаться целесообразным все же оставить в регрессионной модели одну или несколько объясняющих переменных, не оказывающих существенного (значимого) влияния на зависимую переменную.

13.8. Оценка взаимосвязи переменных. Проверка значимости уравнения множественной регрессии

Для оценки взаимосвязи между зависимой переменной и совокупностью объясняющих переменных используется *множественный (совокупный) коэффициент (индекс) корреляции* (см. § 12.6), который может быть выражен через суммы квадратов отклонений по формуле **(13.19)**:

$$R = \sqrt{\frac{Q_R}{Q}} = \sqrt{1 - \frac{Q_e}{Q}},$$

где Q, Q_R и Q_e вычисляются по формулам **(13.16)**, **(13.17)**.

Получим более удобную формулу для R, не требующую вычисления остатков e_i и остаточной суммы квадратов $Q_e = \sum_{i=1}^{n} e_i^2$.

□ В соответствии с **(13.16)**

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

$$= \sum_{i=1}^{n} y_i^2 - 2\bar{y}(n\bar{y}) + n\bar{y}^2 = Y'Y - n\bar{y}^2$$

$$\left(\text{ибо } \sum_{i=1}^{n} y_i^2 = y_1^2 + y_2^2 + \ldots + y_n^2 = (y_1 \;\; y_2 \;\; \ldots \;\; y_n)\begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix} = Y'Y\right).$$

С учетом **(13.24′)** имеем

$$Q_e = \sum_{i=1}^{n} (y_i - y_{x_i})^2 = Y'Y - 2b'X'Y + b'X'Xb = Y'Y - b'X'Y$$

(ибо в силу **(13.25)** $b'X'Xb = b'X'Y$).

Наконец,

$$Q_R = Q - Q_e = Y'Y - n\bar{y}^2 - (Y'Y - b'X'Y) = b'X'Y - n\bar{y}^2.$$

Таким образом,

$$R = \sqrt{\frac{Q_R}{Q}} = \sqrt{\frac{b'X'Y - n\bar{y}^{2'}}{Y'Y - n\bar{y}^2}}. \blacksquare \qquad \textbf{(13.44)}$$

Коэффициент R является обобщением коэффициента корреляции в множественной модели. В зависимости от тесноты связи R может принимать значения от 0 до 1. Величина R^2, называемая *множественным коэффициентом детерминации*, показывает долю вариации зависимой переменной, обусловленную регрессией или изменчивостью объясняющих переменных.

Таким образом, *множественный коэффициент детерминации R^2 можно рассматривать как меру качества уравнения регрессии, характеристику прогностической силы анализируемой регрессионной модели*: чем ближе R^2 к единице, тем лучше регрессия описывает зависимость между объясняющими и зависимой переменными.

Недостатком коэффициента детерминации R^2 является то, что он, вообще говоря, увеличивается при добавлении новых объясняющих переменных, хотя это и не обязательно означает улучшение качества регрессионной модели. В этом смысле предпочтительнее использовать *скорректированный (адаптированный, поправленный) коэффициент детерминации* $\hat{R}^2$, определяемый по формуле

$$\hat{R}^2 = 1 - \frac{n-1}{n-p-1}\left(1-R^2\right). \qquad \textbf{(13.44')}$$

Из **(13.44')** следует, что чем больше число объясняющих переменных p, тем меньше $\hat{R}^2$ по сравнению с R^2. В отличие от R^2 скорректированный коэффициент $\hat{R}^2$ может уменьшаться при введении в модель новых объясняющих переменных, не оказывающих существенного влияния на зависимую переменную. Однако даже увеличение скорректированного коэффициента детерминации $\hat{R}^2$ при введении в модель новой объясняющей переменной не всегда означает, что ее коэффициент регрессии значим (это происходит, как можно показать, только в случае, если соответствующее значение t-статистики больше единицы (по абсолютной величине), т.е. $|t|>1$). Другими словами, увеличение $\hat{R}^2$ еще не означает улучшение качества регрессионной модели.

Если известен коэффициент детерминации R^2, то критерий значимости **(13.18)** уравнения регрессии может быть записан в виде:

$$F = \frac{R^2(n-p-1)}{(1-R^2)p} > F_{\alpha;k_1;k_2}, \qquad \textbf{(13.45)}$$

где $k_1 = p$, $k_2 = n - p - 1$, ибо в уравнении множественной регрессии вместе со свободным членом оценивается $m = p + 1$ параметров.

▷ **Пример 13.7.** По данным примера **13.4** определить множественный коэффициент (индекс) корреляции и проверить на уровне $\alpha = 0{,}05$ значимость полученного уравнения регрессии Y по X_1 и X_2.

Р е ш е н и е. Вычислим произведения векторов (см. пример **13.4**):

$$b'X'Y = \begin{pmatrix}-3{,}54 & 0{,}854 & 0{,}367\end{pmatrix}\begin{pmatrix}68\\664\\445\end{pmatrix} =$$

$$= -3{,}54\cdot 68 + 0{,}854\cdot 664 + 0{,}367\cdot 445 = 489{,}65$$

и $Y'Y = \sum_{i=1}^{10} y_i^2 = 496$ (см. итоговую строку табл. 13.7). Из табл. 13.7 находим также $\sum_{i=1}^{10} y_i = 68$, откуда $\bar{y} = \sum_{i=1}^{n} y_i/n = 68/10 = 6{,}8$ (т).

Теперь по **(13.44)** множественный коэффициент (индекс) корреляции

$$R = \sqrt{\frac{489{,}65 - 10\cdot 6{,}8^2}{496 - 10\cdot 6{,}8^2}} = \sqrt{0{,}811} = 0{,}900.$$

Значение $R = 0{,}900$, близкое к 1, указывает на тесную взаимосвязь зависимой переменной Y — сменной добычи угля на одного рабочего и объясняющих переменных — мощности пласта X_1 и уровня механизации работ X_2. Коэффициент детерминации R^2=0,811 свидетельствует о том, что вариация исследуемой зависимой переменной на 81,1% объясняется изменчивостью включенных в модель объясняющих переменных.

Проделав аналогичные расчеты по данным примера **13.1** для одной объясняющей переменной X_1, можно было получить $R' = 0{,}866$ и $R'^2 = 0{,}751$ (заметим, что в случае одной объясняющей переменной множественный коэффициент корреляции R' равен парному коэффициенту корреляции r). Сравнивая значения R^2 и R'^2, можно сказать, что добавление второй объясняющей переменной X_2 незначительно увеличило коэффициент детерминации, определяющий качество модели. И это понятно, так как выше, в примере **13.6**, мы убедились в незначимости коэффициента регрессии b_2 при переменной X_2.

По формуле **(13.44')** вычислим скорректированный коэффициент детерминации:

при $p = 1$ $\quad \hat{R}'^2 = 1 - \frac{9}{8}(1 - 0{,}751) = 0{,}720;$

при $p = 2$ $\quad \hat{R}'^2 = 1 - \frac{9}{7}(1 - 0{,}811) = 0{,}757.$

Видим, что хотя скорректированный коэффициент детерминации и увеличился при добавлении объясняющей переменной X_2, но это еще не говорит о значимости коэффициента регрессии b_2 (значение t-статистики, равное 1,51 (см. § 13.6), хотя и больше 1, но недостаточно для соответствующего вывода на приемлемом уровне значимости).

Зная R^2=0,811, проверим значимость уравнения регрессии. Фактическое значение критерия по **(13.45)**

$$F = \frac{0{,}811(10-2-1)}{(1-0{,}811)\cdot 2} = 15{,}0$$

больше табличного $F_{0,05;2;7} = 4{,}74$, определенного на уровне значимости $\alpha = 0{,}05$ при $k_1 = 2$ и $k_2 = 10-2-1 = 7$ степенях свободы (см. табл. VI приложений), т.е. уравнение регрессии значимо, следовательно, исследуемая зависимая переменная Y достаточно хорошо описывается включенными в регрессионную модель переменными X_1 и X_2. ▶

13.9. Мультиколлинеарность

*Под **мультиколлинеарностью** понимается высокая взаимная коррелированность объясняющих переменных.* Мультиколлинеарность может проявляться в функциональной (явной) и стохастической (скрытой) формах.

При *функциональной* форме мультиколлинеарности по крайней мере одна из парных связей между объясняющими переменными является линейной функциональной зависимостью. В этом случае матрица $X'X$ особенная, так как содержит линейно зависимые векторы-столбцы и ее определитель равен нулю, т.е. нарушается предпосылка **6** регрессионного анализа. Это приводит к невозможности решения соответствующей системы нормальных уравнений и получения оценок параметров регрессионной модели.

Однако в экономических исследованиях мультиколлинеарность чаще проявляется в *стохастической* форме, когда между хотя бы двумя объясняющими переменными существует тесная корреляционная связь. Матрица $X'X$ в этом случае является неособенной, но ее определитель очень мал. В то же время вектор оценок b и его ковариационная матрица K в соответствии с формулами **(13.28)** и **(13.36)** пропорциональны обратной матрице $(X'X)^{-1}$, а значит, их элементы обратно пропорциональны величине определителя $|X'X|$. В результате получаются значительные средние квадратические отклонения (стандартные ошибки) коэффициентов регрессии $b_0, b_1, \ldots, b_p$ и оценка их значимости по t-критерию не имеет смысла, хотя в целом регрессионная модель может оказаться значимой по F-критерию.

Оценки b_i становятся очень чувствительными к незначительному изменению результатов наблюдений и объема выборки. Уравнения регрессии в этом случае, как правило, не имеют реального смысла, так как некоторые из его коэффициентов могут

иметь неправильные с точки зрения экономической теории знаки и неоправданно большие значения.

Один из методов выявления мультиколлинеарности заключается в анализе корреляционной матрицы между объясняющими переменными $X_1, X_2, \ldots, X_p$ и выявлении пар переменных, имеющих высокие коэффициенты корреляции (обычно больше 0,8). Если такие переменные существуют, то говорят о мультиколлинеарности между ними.

Полезно также находить множественные коэффициенты корреляции между одной из объясняющих переменных и некоторой группой из них. Наличие высокого множественного коэффициента корреляции (обычно принимают больше 0,8) свидетельствует о мультиколлинеарности.

Другой подход состоит в исследовании матрицы $X'X$. Если определитель матрицы $X'X$ близок к нулю (например, одного порядка с накапливающимися ошибками вычислений), то это говорит о наличии мультиколлинеарности.

Для устранения или уменьшения мультиколлинеарности используется ряд методов. Один из них заключается в том, что из двух объясняющих переменных, имеющих высокий коэффициент корреляции (больше 0,8), одну переменную исключают из рассмотрения. При этом, какую переменную оставить, а какую удалить из анализа, решают в первую очередь на основании экономических соображений. Если с экономической точки зрения ни одной из переменных нельзя отдать предпочтение, то оставляют ту из двух переменных, которая имеет больший коэффициент корреляции с зависимой переменной.

Другим из возможных методов устранения или уменьшения мультиколлинеарности является использование пошаговых процедур отбора наиболее информативных переменных. Например, вначале рассматривается линейная регрессия зависимой переменной Y от объясняющей переменной, имеющей с ней наиболее высокий коэффициент корреляции (или индекс корреляции при нелинейной форме связи). На втором шаге включается в рассмотрение та объясняющая переменная, которая имеет наиболее высокий частный коэффициент корреляции с Y и вычисляется множественный коэффициент (индекс) корреляции. На третьем шаге вводится новая объясняющая переменная, которая имеет наибольший частный коэффициент корреляции с Y, и вновь вычисляется множественный коэффициент корреляции и т.д.

Процедура введения новых переменных продолжается до тех пор, пока добавление следующей объясняющей переменной существенно не увеличивает множественный коэффициент корреляции.

13.10. Понятие о других методах многомерного статистического анализа

Многомерный статистический анализ *определяется*[1] *как раздел математической статистики, посвященный математическим методам построения оптимальных планов сбора, систематизации и обработки многомерных статистических данных, направленных на выявление характера и структуры взаимосвязей между компонентами исследуемого признака и предназначенных для получения научных и практических выводов.*

Многомерные статистические методы среди множества возможных вероятностно-статистических моделей позволяют обоснованно выбрать ту, которая наилучшим образом соответствует исходным статистическим данным, характеризующим реальное поведение исследуемой совокупности объектов, оценить надежность и точность выводов, сделанных на основании ограниченного статистического материала.

С некоторыми разделами многомерного статистического анализа, такими, как многомерный корреляционный анализ, множественная регрессия, многомерный дисперсионный анализ (на примере двухфакторного анализа) мы уже сталкивались в гл. 11—13. Приведем теперь краткий обзор ряда других методов многомерного статистического анализа, которые уже нашли отражение в статистических пакетах прикладных программ. В первую очередь следует выделить *методы, позволяющие выявить общие (скрытые или латентные) факторы, определяющие вариацию первоначальных факторов.* К ним относятся факторный анализ и метод главных компонент.

Факторный анализ. Основной задачей факторного анализа является переход от первоначальной системы большого числа взаимосвязанных факторов X_1, X_2,...,X_m к относительно малому числу скрытых (латентных) факторов F_1,F_2,...,F_k, $k < m$. Скажем, производительность труда на предприятиях зависит от множества факторов (образовательного уровня сотрудников, коэффициента сменности оборудования, электровооруженности труда,

[1] См.: *Математическая* энциклопедия, т. 3. — М.: Советская энциклопедия, 1982.

возраста оборудования, количества мест в столовых и т.п.), из которых многие факторы связаны между собой. Используя факторный анализ, можно установить влияние на рост производительности труда лишь нескольких обобщенных факторов (например, размера предприятия, уровня организации труда, характера продукции), непосредственно не наблюдавшихся.

Модель факторного анализа записывается в виде:

$$X_i = a_i + \sum_{j=1}^{k} a_{ij} F_j + v_j \varepsilon_j, \quad i = 1,2,...,m,\ k < m, \qquad \textbf{(13.46)}$$

где $a_i = M(X_i)$ — математическое ожидание первоначального фактора X_i;

F_j — общие (скрытые или латентные) факторы ($j = 1,2,...,k$);

a_{ij} — нагрузки первоначальных факторов на общие факторы;

ε_i — характерные факторы ($i = 1,2,...,m$);

v_j — нагрузки первоначальных факторов на характерные факторы.

Первое слагаемое в модели **(13.46)** — неслучайная составляющая, другие два слагаемых — случайные составляющие.

Особенностью факторного анализа является неоднозначность определения общих факторов.

Метод главных компонент (компонентный анализ). В отличие от рассматриваемых в факторном анализе общих факторов, которые обусловливают большую (но не всю) часть вариации первоначальных факторов, главные компоненты объясняют всю вариацию и определяются однозначно. Модель главных компонент имеет вид:

$$X_i = a_i + \sum_{j=1}^{m} a_{ij} F_j, \quad i = 1,2,...,m\ . \qquad \textbf{(13.47)}$$

Как видим, в модели **(13.47)** отсутствуют характерные факторы, так как главные компоненты F_j полностью обусловливают всю вариацию первоначальных факторов.

Для углубления анализа изучаемого явления после выявления главных компонент рассматривают регрессию на главных компонентах, в которых последние выступают в качестве обобщенных объясняющих переменных.

Среди других методов многомерного статистического анализа отметим *методы, позволяющие осуществить классификацию экономических объектов*, т.е. отнесение их к определенным классам. Это методы дискриминантного и кластерного анализа.

Дискриминантный анализ позволяет отнести объект, характеризующийся значениями m признаков, к одной из l совокупностей (классов, групп), заданных своими распределениями. Предполагается, что l совокупностей заданы выборками (называемыми *обучаемыми*), которые содержат информацию о статистических распределениях совокупностей в m-мерном пространстве признаков.

При отсутствии обучающих выборок могут быть использованы методы **кластерного анализа**, позволяющие разбить исследуемую совокупность объектов на группы «схожих» объектов, называемых *кластерами*, таким образом, чтобы объекты одного класса находились на «близких» расстояниях между собой, а объекты разных классов — на относительно «отдаленных» расстояниях друг от друга. При этом каждый объект X_j ($j = 1,2,...,m$) рассматривается как точка в m-мерном пространстве, и выбор способа вычисления расстояний или близости между объектами и признаками является узловым моментом исследования, от которого в основном зависит окончательный вариант разбиения объектов на классы.

В завершение краткого обзора отметим, что применение методов многомерного статистического анализа невозможно без использования пакетов прикладных программ (см., например, [30]). Подробное изложение многомерных статистических методов приведено, в частности, в учебнике [13].

Упражнения

13.8. По данным примера **12.14**: а) найти уравнение регрессии Y по X; б) оценить среднюю энерговооруженность труда на предприятиях, фондовооруженность которых равна 10 млн руб. , и построить для нее 95%-ный доверительный интервал; в) найти коэффициент детерминации R^2 и пояснить его смысл; г) проверить значимость уравнения регрессии на 5%-ном уровне по F-критерию.

13.9. По данным примера **12.15**: а) найти уравнение регрессии Y по X; б) найти коэффициент детерминации r^2 и пояснить его смысл; в) проверить значимость уравнения регрессии на 5%-ном уровне по F-критерию; г) оценить среднюю производительность труда на предприятиях с уровнем механизации работ 60% и построить для нее 95%-ный доверительный интервал; аналогичный доверительный интервал найти для индивидуальных значений производительности труда на тех же предприятиях.

13.10. По данным 30 нефтяных компаний получено следующее уравнение регрессии между оценкой Y (ден. ед.) и фактической стоимостью X (ден. ед.) этих компаний: $y_x = 0{,}8750x + 295$. Найти: 95%-ные доверительные интервалы для среднего и индивидуального значений оценки предприятий, фактическая стоимость которых составила 1300 ден. ед., если коэффициент корреляции между переменными равен 0,76, среднее значение переменной X равно 1430 ден. ед., а ее среднее квадратическое отклонение равно 270 ден. ед.

13.11. На 10 опытных участках одинакового размера получены следующие данные об урожайности X (т) и содержании белка Y (%) для некоторой культуры:

Урожайность, т	9,9	10,2	11,0	11,6	11,8	12,5	12,8	13,5	14,3	14,4
Содержание белка, %	10,7	10,8	12,1	12,5	12,8	12,8	12,4	11,8	10,8	10,6

Необходимо: а) выровнять зависимость Y от X по параболе второго порядка и проверить значимость полученного уравнения регрессии; б) оценить тесноту связи между переменными с помощью индекса корреляции R_{yx} и коэффициента детерминации R^2_{yx}; в) определить, при каком значении урожайности средний процент содержания белка будет максимальным; найти этот процент.

13.12. Распределение 50 гастрономических магазинов области по уровню издержек обращения X (%) и годовому объему товарооборота Y (млн руб.) представлено в таблице:

x \ y	0,5—2,0	2,0—3,5	3,5—5,0	5,0—6,5	6,5—8,0	Итого
4—6	—	—	—	3	2	5
6—8	—	4	8	8	1	21
8—10	2	5	5	2	—	14
10—12	3	1	5	—	—	9
12—14	1	—	—	—	—	1
Итого	6	10	18	13	3	50

Необходимо: а) построить эмпирическую линию регрессии Y по X; б) выровнять полученную зависимость по прямой и гиперболе и вычислить остаточную дисперсию для каждого случая; в) оценить тесноту связи между переменными с помощью эмпирического корреляционного отношения η_{yx}, коэффициента корреляции r и индекса корреляции R_{yx}; проверить значимость η_{yx} и R_{yx} и сравнить их по величине; г) на основании результатов, полученных в пп. а), б), в), определить, какое из двух полученных уравнений регрессии целесообразнее использовать для исследования заданной зависимости.

13.13. Имеются следующие данные о выработке литья на одного работающего X_1 (т), браке литья X_2 (%) и себестоимости одной тонны литья Y (руб.) по 25 литейным цехам заводов:

i	x_{1i}	x_{2i}	y_i	i	x_{1i}	x_{2i}	y_i	i	x_{1i}	x_{2i}	y_i
1	14,6	4,2	239	10	25,3	0,9	198	19	17,0	9,3	282
2	13,5	6,7	254	11	56,0	1,3	170	20	33,1	3,3	196
3	21,5	5,5	262	12	40,2	1,8	173	21	30,1	3,5	186
4	17,4	7,7	251	13	40,6	3,3	197	22	65,2	1,0	176
5	44,8	1,2	158	14	75,8	3,4	172	23	22,6	5,2	238
6	111,9	2,2	101	15	27,6	1,1	201	24	33,4	2,3	204
7	20,1	8,4	259	16	88,4	0,1	130	25	19,7	2,7	205
8	28,1	1,4	186	17	16,6	4,1	251				
9	22,3	4,2	204	18	33,4	2,3	195				

Необходимо: а) найти парные, частные и множественный $R_{y.12}$ коэффициенты корреляции между переменными и оценить их значимость на уровне $\alpha = 0{,}05$; б) найти уравнение множественной регрессии Y по X_1 и X_2, оценить значимость этого уравнения и его коэффициентов на уровне $\alpha = 0{,}05$; в) сравнить раздельное влияние на зависимую переменную каждой из объясняющих переменных, используя стандартизованные коэффициенты регрессии и коэффициенты эластичности; г) найти 95%-ные доверительные интервалы для

коэффициентов регрессии, а также для среднего и индивидуальных значений себестоимости одной тонны литья в цехах, в которых выработка литья на одного работающего составляет 40 т, а брак литья — 5%.

13.14. Имеются следующие данные о годовых ставках месячных доходов по трем акциям за шестимесячный период:

Акция	*Доходы по месяцам, %*					
A	5,4	5,3	4,9	4,9	5,4	6,0
B	6,3	6,2	6,1	5,8	5,7	5,7
C	9,2	9,2	9,1	9,0	8,7	8,6

Есть основания предполагать, что доходы Y по акции C зависят от доходов X_1 и X_2 по акциям A и B. Необходимо: а) составить уравнение регрессии Y по X_1 и X_2; б) найти множественный коэффициент корреляции R и коэффициент детерминации R^2 и пояснить их смысл; в) проверить на уровне $\alpha = 0{,}05$ значимость полученного уравнения регрессии; г) оценить средний доход по акции C, если доходы по акциям A и B составили соответственно 5,5 и 6,0%.

Глава 14 Введение в анализ временных рядов

14.1. Общие сведения о временных рядах и задачах их анализа

Анализ временных рядов представляет собой самостоятельную, весьма обширную и одну из наиболее интенсивно развивающихся областей математической статистики.

Под ***временным рядом*** (***динамическим рядом***, или ***рядом динамики***) в экономике подразумевается последовательность наблюдений некоторого признака (случайной величины) X в последовательные равноотстоящие моменты времени. Отдельные наблюдения называются *уровнями* ряда, которые будем обозначать $x_t\,(t=1,2,...,n)$, где n — число уровней.

В табл. 14.1 приведены данные, отражающие цену и спрос на некоторый товар за восьмилетний период (усл. ед.), т.е. два временных ряда — цены товара x_t и спроса y_t на него.

Таблица 14.1

Год, t	1	2	3	4	5	6	7	8
Цена, x_t	492	462	350	317	340	351	368	381
Спрос, y_t	213	171	291	309	317	362	351	361

В качестве примера на рис. 14.1 временной ряд y_t изображен графически.

В общем виде при исследовании экономического временного ряда x_t выделяются несколько составляющих:

$$x_t = u_t + v_t + c_t + \varepsilon_t \quad (t = 1,2,...,n),$$

где u_t — *тренд*, плавно меняющаяся компонента, описывающая чистое влияние долговременных факторов, т.е. длительную («вековую») тенденцию изменения признака (например, рост населения, экономическое развитие, изменение структуры потребления и т.п.);

v_t — *сезонная компонента*, отражающая повторяемость экономических процессов в течение не очень длительного периода (года, иногда месяца, недели и т.д., например, объем продаж товаров или перевозок пассажиров в различные времена года);

c_t — *циклическая компонента*, отражающая повторяемость экономических процессов в течение длительных периодов (например, влияние волн экономической активности Кондратьева, демографических «ям», циклов солнечной активности и т.п.);

ε_t — *случайная компонента*, отражающая влияние не поддающихся учету и регистрации случайных факторов.

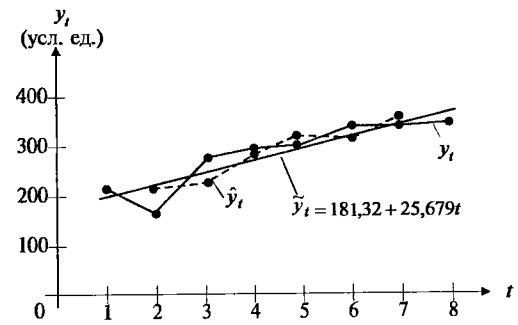


Рис. 14.1

Следует обратить внимание на то, что в отличие от ε_t первые три составляющие (компоненты) u_t, v_t, c_t являются *закономерными, неслучайными.*

*Важнейшей классической задачей при исследовании экономических временных рядов является выявление и статистическая оценка **основной тенденции** развития изучаемого процесса и отклонений от нее.*

Отметим основные этапы анализа временных рядов:

- графическое представление и описание поведения временного ряда;
- выделение и удаление закономерных (неслучайных) составляющих временного ряда (тренда, сезонных и циклических составляющих);
- сглаживание и фильтрация (удаление низко- или высокочастотных составляющих временного ряда);
- исследование случайной составляющей временного ряда, построение и проверка адекватности математической модели для ее описания;
- прогнозирование развития изучаемого процесса на основе имеющегося временного ряда;
- исследование взаимосвязи между различными временными рядами.

Среди наиболее распространенных методов анализа временных рядов выделим *корреляционный* и *спектральный анализ, модели авторегрессии* и *скользящей средней.* О некоторых из них речь пойдет ниже.

Так же, как ранее вариационный ряд $x_1,x_2,...,x_i,...,x_n$ рассматривался как одна из реализаций случайной величины X, временной ряд $x_1, x_2,...,x_i,...,x_n$ рассматривается как одна из *реализаций* (*траекторий*) случайного процесса $X(t)$ (см. § 7.1). Вместе с тем следует иметь в виду принципиальные отличия временного ряда x_t ($t = 1,2,...,n$) от последовательности наблюдений x_1, $x_2,...,x_n$, образующих случайную выборку. Во первых, в отличие от элементов выборки члены временного ряда, как правило, не являются статистически независимыми. Во-вторых, члены временного ряда не являются одинаково распределенными.

14.2. Стационарные временные ряды и их характеристики. Автокорреляционная функция

Важное значение в анализе временных рядов имеют стационарные временные ряды, вероятностные свойства которых не изменяются во времени. Стационарные временные ряды применяются, в частности, при описании случайных составляющих анализируемых рядов.

Временной ряд x_t ($t = 1,2,...,n$) называется *строго стационарным* (или *стационарным в узком смысле*), если совместное распределение вероятностей n наблюдений x_1, $x_2,...,x_n$ такое же, как и n наблюдений $x_{1+\tau}$, $x_{2+\tau},...,x_{n+\tau}$ при любых n, t и τ. Другими словами, свойства строго стационарных рядов x_t не зависят от момента t, т.е. закон распределения и его числовые характеристики не зависят от t. Следовательно, математическое ожидание $a_x(t) = a$, среднее квадратическое отклонение $\sigma_x(t) = \sigma$ (см. § 7.1) могут быть оценены по наблюдениям x_t ($t = 1,2,...,n$) по формулам:

$$\overline{x}_t = \frac{\sum_{t=1}^{n} x_t}{n}, \qquad \textbf{(14.1)}$$

$$s_t^2 = \frac{\sum_{t=1}^{n} (x_t - \overline{x}_t)^2}{n}. \qquad \textbf{(14.2)}$$

Степень тесноты связи между последовательностями наблюдений временного ряда x_1, $x_2,...,x_n$ и $x_{1+\tau}$, $x_{2+\tau},...,x_{n+\tau}$ (сдвинутых относительно друг друга на τ единиц, или, как говорят, с *лагом* τ) может быть определена с помощью коэффициента корреляции

$$\rho(\tau) = \frac{M\left[(x_t - a)(x_{t+\tau} - a)\right]}{\sigma_x(t)\sigma_x(t+\tau)} = \frac{M\left[(x_t - a)(x_{t+\tau} - a)\right]}{\sigma^2}, \qquad \textbf{(14.3)}$$

ибо $M(x_t) = M(x_{t+\tau}) = a$, $\sigma_x(t) = \sigma_x(t+\tau) = \sigma$.

Так как коэффициент $\rho(\tau)$ измеряет корреляцию между членами одного и того же ряда, его называют *коэффициентом автокорреляции*, а зависимость $\rho(\tau)$ — *автокорреляционной функцией.* В силу стационарности временного ряда x_t ($t = 1,2,...,n$) автокорреляционная функция $\rho(\tau)$ зависит только от лага τ, причем $\rho(-\tau) = \rho(\tau)$, т.е. при изучении $\rho(\tau)$ можно ограничиться рассмотрением только положительных значений τ.

Статистической оценкой $\rho(\tau)$ является *выборочный коэффициент автокорреляции* r_τ, определяемый по формуле коэффициента корреляции **(12.35')**, в которой $x_i = x_t$, $y_i = x_{t+\tau}$, а n заменяется на $n-\tau$:

$$r_\tau = \frac{(n-\tau)\sum_{t=1}^{n-\tau} x_t x_{t+\tau} - \sum_{t=1}^{n-\tau} x_t \sum_{t=1}^{n-\tau} x_{t+\tau}}{\sqrt{(n-\tau)\sum_{t=1}^{n-\tau} x_t^2 - \left(\sum_{t=1}^{n-\tau} x_t\right)^2}\sqrt{(n-\tau)\sum_{t=1}^{n-\tau} x_{t+\tau}^2 - \left(\sum_{t=1}^{n-\tau} x_{t+\tau}\right)^2}}. \qquad \textbf{(14.4)}$$

Функцию r_τ называют *выборочной автокорреляционной функцией*, а ее график — *коррелограммой.*

При расчете r_τ следует помнить, что с увеличением τ число $n-\tau$ пар наблюдений x_t, $x_{t+\tau}$ уменьшается, поэтому лаг τ должен быть таким, чтобы число $n-\tau$ было достаточным для определения r_τ. Обычно ориентируются на соотношение $\tau \le n/4$.

Для стационарного временного ряда с увеличением лага τ взаимосвязь членов временного ряда x_t и $x_{t+\tau}$ ослабевает и автокорреляционная функция $\rho(\tau)$ должна убывать (по аб-

солютной величине). В то же время для ее выборочного (эмпирического) аналога r_τ, особенно при небольшом числе пар наблюдений $n-\tau$, свойство монотонного убывания (по абсолютной величине) при возрастании τ может нарушаться.

▷ **Пример 14.1.** По данным табл. 14.1 для временного ряда y_t найти среднее значение, среднее квадратическое отклонение и коэффициенты автокорреляции (для лагов $\tau = 1;2$).

Р е ш е н и е. Среднее значение временного ряда находим по формуле (**14.1**):

$$\bar{y}_t = \frac{213+171+...+361}{8} = 296{,}88\ (\text{ед.}).$$

Дисперсию и среднее квадратическое отклонение можно вычислить по формуле (**14.2**), но в данном случае проще использовать соотношение

$$s_t^2 = \overline{y_t^2} - \bar{y}_t^2 = 92\,478{,}38 - 296{,}88^2 = 4343{,}61;$$

$$s_t = \sqrt{4343{,}61} = 65{,}31\ (\text{ед.}),$$

где $$\overline{y_t^2} = \frac{\sum_{t=1}^{n} y_t^2}{n} = \frac{213^2+171^2+...+361^2}{8} = 92\,478{,}38.$$

Найдем коэффициент автокорреляции r_τ временного ряда (для лага $\tau = 1$), т.е. коэффициент корреляции между последовательностями семи пар наблюдений y_t и y_{t+1} ($t = 1,2,...,7$):

y_t	213	171	291	309	317	362	351
$y_{t+\tau}$	171	291	309	317	362	351	361

Вычисляем необходимые суммы:

$$\sum_{t=1}^{7} y_t = 213+171+...+351 = 2014;$$

$$\sum_{t=1}^{7} y_t^2 = 213^2+171^2+...+351^2 = 609\,506;$$

$$\sum_{t=1}^{7} y_{t+\tau} = 171+291+...+361 = 2162;$$

$$\sum_{t=1}^{7} y_{t+\tau}^2 = 171^2+291^2+...+361^2 = 694\,458;$$

$$\sum_{t=1}^{7} y_t y_{t+\tau} = 213\cdot 171+171\cdot 291+...+351\cdot 361 = 642\,583.$$

Теперь по формуле (**14.4**) коэффициент автокорреляции

$$r_1 = \frac{7\cdot 642\,583 - 2014\cdot 2162}{\sqrt{7\cdot 609\,506 - 2014^2}\sqrt{7\cdot 694\,458 - 2162^2}} = 0{,}725.$$

Вычисление коэффициента автокорреляции r_2 временного ряда $y(t)$ для лага $\tau=2$, т.е. коэффициента корреляции между последовательностями шести пар наблюдений y_t и y_{t+2} предлагаем провести читателю самостоятельно. ▶

Знание автокорреляционной функции r_τ может оказать существенную помощь при подборе модели анализируемого временного ряда и статистической оценке ее параметров.

14.3. Аналитическое выравнивание (сглаживание) временного ряда (выделение неслучайной компоненты)

Как уже отмечено выше, одной из важнейших задач исследования экономического временного ряда является выявление *основной тенденции* изучаемого процесса, выраженной неслучайной составляющей $f(t)$ (тренда либо тренда с циклической или (и) сезонной компонентой).

Для решения этой задачи вначале необходимо выбрать в и д функции $f(t)$. Наиболее часто используются следующие функции:

линейная — $f(t) = b_0 + b_1 t$;

полиномиальная — $f(t) = b_0 + b_1 t + b_2 t^2 + ... + b_n t^n$;

экспоненциальная — $f(t) = e^{b_0 + b_1 t}$;

логистическая — $f(t) = \dfrac{a}{1 + be^{-ct}}$;

Гомперца — $\log_c f(t) = a - br^t$, где $0 < r < 1$.

Это весьма ответственный этап исследования. При выборе соответствующей функции $f(t)$ используют содержательный ана-

лиз (который может установить характер динамики процесса), визуальные наблюдения (на основе графического изображения временного ряда). При выборе полиномиальной функции может быть применен метод последовательных разностей (состоящий в вычислении разностей первого порядка $\Delta_t = x_t - x_{t-1}$, второго порядка $\Delta_t^{(2)} = \Delta_t - \Delta_{t-1}$ и т.д., и порядок разностей, при котором они будут примерно одинаковыми, принимается за степень полинома).

Из двух функций предпочтение обычно отдается той, при которой меньше сумма квадратов отклонений фактических данных от расчетных на основе этих функций. Но этот принцип нельзя доводить до абсурда: так, для любого ряда из n точек можно подобрать полином $(n-1)$-й степени, проходящий через все точки, и соответственно с минимальной — нулевой — суммой квадратов отклонений, но в этом случае, очевидно, не следует говорить о выделении основной тенденции, учитывая случайный характер этих точек. Поэтому при прочих равных условиях предпочтение следует отдавать более простым функциям. (Подробнее об этом см. § 13.3.)

Для выявления основной тенденции чаще всего используется **метод наименьших квадратов**, рассмотренный в гл. 12. Значения временного ряда x_t или y_t рассматриваются как зависимая переменная, а время t — как объясняющая:

$$y_t = f(t) + \varepsilon_t, \qquad \textbf{(14.5)}$$

где ε_t — возмущения, удовлетворяющие основным предпосылкам регрессионного анализа, приведенным в § 13.1, т.е. представляющие независимые и одинаково распределенные случайные величины, распределение которых предполагаем нормальным.

Напомним, что согласно методу наименьших квадратов параметры прямой $\tilde{y}_t = b_0 + b_1 t$ находятся из системы нормальных уравнений (**12.10**), в которой в качестве x_i берем t, а $n_i = 1$:

$$\begin{cases} b_0 n + b_1 \sum\limits_{t=1}^{n} t = \sum\limits_{t=1}^{n} y_t, \\ b_0 \sum\limits_{t=1}^{n} t + b_1 \sum\limits_{t=1}^{n} t^2 = \sum\limits_{t=1}^{n} y_t t, \end{cases}$$

а параметры параболы $\tilde{y}_t = b_0 + b_1 t + b_2 t^2$ — из системы нормальных уравнений (**13.21**):

$$\begin{cases} b_0 n + b_1 \sum\limits_{t=1}^{n} t + b_2 \sum\limits_{t=1}^{n} t^2 = \sum\limits_{t=1}^{n} y_t, \\ b_0 \sum\limits_{t=1}^{n} t + b_1 \sum\limits_{t=1}^{n} t^2 + b_2 \sum\limits_{t=1}^{n} t^3 = \sum\limits_{t=1}^{n} y_t t, \\ b_0 \sum\limits_{t=1}^{n} t^2 + b_1 \sum\limits_{t=1}^{n} t^3 + b_2 \sum\limits_{t=1}^{n} t^4 = \sum\limits_{t=1}^{n} y_t t^2. \end{cases}$$

Учитывая, что значения переменной $t = 1,2,...,n$ образуют натуральный ряд чисел от 1 до n, суммы $\sum\limits_{t=1}^{n} t$, $\sum\limits_{t=1}^{n} t^2$, $\sum\limits_{t=1}^{n} t^3$, $\sum\limits_{t=1}^{n} t^4$ можно выразить через число членов ряда n по известным в математике формулам:

$$\sum_{t=1}^{n} t = \frac{n(n+1)}{2}; \quad \sum_{t=1}^{n} t^2 = \frac{n(n+1)(2n+1)}{6}; \qquad \textbf{(14.6)}$$

$$\sum_{t=1}^{n} t^3 = \frac{n^2(n+1)^2}{4}; \quad \sum_{t=1}^{n} t^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}. \qquad \textbf{(14.7)}$$

▷ **Пример 14.2.** По данным табл. 14.1 найти уравнение неслучайной составляющей (тренда) для временного ряда y_t, полагая тренд линейным.

Р е ш е н и е. По формуле (**14.6**):

$$\sum_{t=1}^{8} t = \frac{8 \cdot 9}{1 \cdot 2} = 36; \quad \sum_{t=1}^{8} t^2 = \frac{8 \cdot 9 \cdot 17}{6} = 204;$$

Далее $\sum\limits_{t=1}^{8} y_t = 213+171+...+361=2375$; $\sum\limits_{t=1}^{8} y_t^2 = 213^2+171^2+...+361^2 = 739\,827$; $\sum\limits_{t=1}^{8} y_t t = 213 \cdot 1+171 \cdot 2+...+361 \cdot 8=11\,766$

Система нормальных уравнений имеет вид:

$$\begin{cases} 8b_0 + 36b_1 = 2375, \\ 36b_0 + 204b_1 = 11\,766, \end{cases}$$

откуда $b_0 = 181{,}32$; $b_1 = 25{,}679$ и уравнение тренда $\tilde{y}_t = 181{,}32 + 25{,}679t$ (см. рис. 14.1), т.е. спрос ежегодно увеличивается в среднем на 25,7 ед.

При решении задачи можно было не выписывать систему нормальных уравнений, а представить уравнение регрессии в виде (**12.15**), т.е. $\tilde{y}_t - \bar{y} = b_1(t - \bar{t})$, где $\bar{t} = \dfrac{\sum_{t=1}^{n} t}{n} = \dfrac{1+n}{2}$, $\bar{y} = \dfrac{\sum_{t=1}^{n} y_t}{n}$, а коэффициент регрессии b_1 найти по формуле (**12.17**):

$$b_1 = \frac{\overline{y_t t} - \bar{y}_t \bar{t}}{\overline{t^2} - \bar{t}^2},$$

где $\overline{y_t t} = \sum_{t=1}^{n} y_t t / n$; $\bar{y}_t = \sum_{t=1}^{n} y_t / n$; $\bar{t} = (1+n)/2$; $\overline{t^2} = (n+1)(2n+1)/6$.

Проверим значимость полученного уравнения тренда по F-критерию на 5%-ном уровне значимости. Вычислим с помощью формул (**13.16**), (**13.17′**) суммы квадратов:

а) обусловленную регрессией —

$$Q_R = \sum_{t=1}^{n} (\tilde{y}_t - \bar{y}_t)^2 = \sum_{t=1}^{n} b_1^2 (t - \bar{t})^2 = b_1^2 \left(\sum_{t=1}^{n} t^2 - \frac{\left(\sum_{t=1}^{n} t \right)^2}{n} \right) =$$

$$= 25{,}679^2 \left(204 - \frac{36^2}{8} \right) = 27\,695{,}3;$$

б) общую —

$$Q = \sum_{t=1}^{n} (y_t - \bar{y}_t)^2 = \sum_{t=1}^{n} y_t^2 - \frac{\left(\sum_{t=1}^{n} y_t \right)^2}{n} = 739\,827 - \frac{2375^2}{8} = 34\,748{,}9;$$

в) остаточную —

$$Q_e = Q - Q_R = 34\,748{,}9 - 27\,695{,}3 = 7053{,}6.$$

Найдем по формуле (**13.18′**) значение статистики

$$F = \frac{Q_R (n-2)}{Q_e} = \frac{27\,695{,}3 \cdot 6}{7053{,}6} = 23{,}56.$$

Так как $F > F_{0,05;1;6} = 5{,}99$ (см. табл. VI приложений), то уравнение тренда значимо. ▶

При применении метода наименьших квадратов для оценки параметров экспоненциальной, логистической функций или функции Гомперца возникают сложности с решением получаемой системы нормальных уравнений, поэтому предварительно, до получения соответствующей системы, прибегают к некоторым преобразованиям этих функций (например, логарифмированию и др.).

Другим методом выравнивания (сглаживания) временного ряда, т.е. выделения неслучайной составляющей, является **метод скользящих средних**. Он основан на переходе от начальных значений членов ряда к их средним значениям на интервале времени, длина которого определена заранее. При этом сам выбранный интервал времени «скользит» вдоль ряда.

Получаемый таким образом ряд скользящих средних ведет себя более гладко, чем исходный ряд, из-за усреднения отклонений ряда. Действительно, если индивидуальный разброс значений члена временного ряда x_t около своего среднего (сглаженного) значения a характеризуется дисперсией σ^2, то разброс средней из m членов временного ряда $(x_1 + x_2 + \ldots + x_m)/m$ около того же значения a будет характеризоваться существенно меньшей величиной дисперсии, равной σ^2/m. Для усреднения могут быть использованы средняя арифметическая (простая и с некоторыми весами), медиана и др.

▷ **Пример 14.3.** Провести сглаживание временного ряда y_t по данным табл. 14.1 методом скользящих средних, используя простую среднюю арифметическую с интервалом сглаживания $m = 3$ года.

Р е ш е н и е. Скользящие средние находим по формуле:

$$\hat{y}_t = \frac{\sum_{i=t-p}^{t+p} y_i}{m}, \qquad (14.8)$$

когда $m = (2p-1)$ — нечетное число; при $m = 3$ $p = 1$.

Например, при $t = 2$ по формуле (**14.8**):

$$\hat{y}_2 = \frac{1}{3}(y_1 + y_2 + y_3) = \frac{1}{3}(213 + 171 + 291) = 225 \text{ (ед.)};$$

при $t = 3$ $\hat{y}_3 = \frac{1}{3}(y_2 + y_3 + y_4) = \frac{1}{3}(171 + 291 + 309) = 257{,}0 \text{(ед.)}$ и т.д.

В результате получим сглаженный ряд:

t	1	2	3	4	5	6	7	8
y_t	—	225,0	257,0	305,7	329,3	343,3	358,0	—

На рис. 14.1 этот ряд изображен графически в виде пунктирной линии. ▶

14.4. Временные ряды и прогнозирование. Автокорреляция возмущений

Одна из важнейших задач (этапов) анализа временного (динамического) ряда, как отмечено выше, состоит в прогнозировании на его основе развития изучаемого процесса. При этом исходят из того, что тенденция развития, установленная в прошлом, может быть распространена (экстраполирована) на будущий период.

Задача ставится так: имеется временной (динамический) ряд $y_t(t=1,2,...,n)$ и требуется дать прогноз уровня этого ряда на момент $n+\tau$.

Выше, в § 13.2, 13.6, 13.7, мы рассматривали *точечный и интервальный прогнозы* значений зависимой переменной Y, т.е. *определение точечных и интервальных оценок* Y, полученных для парной и множественной регрессий для значений объясняющих переменных X, расположенных вне пределов обследованного диапазона значений X.

Если рассматривать временной ряд как регрессионную модель изучаемого признака по переменной «время», то к нему могут быть применены рассмотренные выше методы анализа. Следует, однако, вспомнить, что одна из основных предпосылок регрессионного анализа состоит в том, что возмущения $\varepsilon_t(t=1,2,...,n)$ представляют собой независимые случайные величины с математическим ожиданием (средним значением), равным нулю. А при работе с временными рядами такое допущение оказывается во многих случаях неверным.

Действительно, если вид функции тренда выбран неудачно, то вряд ли можно говорить о том, что отклонения от нее (возмущения ε_t) являются независимыми. В этом случае наблюдается заметная концентрация положительных и отрицательных возмущений, и можно предполагать их взаимосвязь. Если последовательные значения ε_t коррелируют между собой, то говорят об *автокорреляции возмущений* (*остатков*, *ошибок*).

Метод наименьших квадратов, вообще говоря, и в случае автокорреляции возмущений дает несмещенные и состоятельные оценки параметров, однако их интервальные оценки могут содержать грубые ошибки. В случае выявления автокорреляции возмущений целесообразно вновь вернуться к проблеме спецификации уравнения регрессии (выбора функции тренда), пересмотреть набор включенных в него переменных и т.п.

Наиболее простым и достаточно надежным критерием определения автокорреляции возмущений является *критерий Дарбина—Уотсона*. С помощью этого критерия проверяется гипотеза об отсутствии автокорреляции между соседними остаточными членами ряда e_t и e_{t-1} (для лага $\tau = 1$), где e_t — выборочная оценка ε_t.

Статистика критерия имеет вид:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad \textbf{(14.9)}$$

При достаточно большом n можно считать, что $\sum_{t=1}^{n} e_t^2 \approx \sum_{t=2}^{n} e_t^2 \approx \sum_{t=2}^{n} e_{t-1}^2$. Тогда после несложных преобразований получим, что

$$d \approx \frac{2\sum_{t=2}^{n} e_t^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \approx 2\left(1 - \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}\right). \qquad \textbf{(14.10)}$$

Статистика d заключена в границах от 0 до 4; при отсутствии автокорреляции $d \approx 2$ (так как при этом $\sum_{t=2}^{n} e_t e_{t-1} = 0$); при полной положительной автокорреляции $d \approx 0\left(\sum_{t=2}^{n} e_t e_{t-1} \approx \sum_{t=2}^{n} e_t^2\right)$; при полной отрицательной — $d \approx 4\left(\sum_{t=2}^{n} e_t e_{t-1} \approx -\sum_{t=2}^{n} e_t^2\right)$.

Для d-статистики найдены верхняя $d_в$ и нижняя $d_н$ критические границы на уровнях значимости $\alpha = 0{,}01; 0{,}025$ и $0{,}05$.

Если фактически наблюдаемое значение d:

а) $d_в < d < 4 - d_в$, то гипотеза об отсутствии автокорреляции не отвергается (принимается);

б) $d_н \le d \le d_в$ или $4-d_в \le d \le 4-d_н$, то вопрос об отвержении или принятии гипотезы остается открытым (область неопределенности критерия);

в) $0<d<d_н$, то принимается альтернативная гипотеза о положительной автокорреляции;

г) $4-d_н<d<4$, то принимается альтернативная гипотеза об отрицательной автокорреляции.

В табл. 14.2 приведен фрагмент таблицы значений статистик $d_н$ и $d_в$ критерия Дарбина—Уотсона на уровне значимости $\alpha = 0{,}05$.

Таблица 14.2

Число наблюдений n	*Число объясняющих переменных*					
	p=1		*p*=2		*p*=3	
	$d_н$	$d_в$	$d_н$	$d_в$	$d_н$	$d_в$
15	1,08	1,36	0,95	1,54	0,82	1,75
16	1,10	1,37	0,98	1,54	0,86	1,73
17	1,13	1,38	1,02	1,54	0,90	1,71
18	1,16	1,39	1,05	1,53	0,93	1,69
19	1,18	1,40	1,08	1,53	0,97	1,68
20	1,20	1,41	1,10	1,54	1,00	1,68
25	1,29	1,45	1,21	1,55	1,12	1,66
30	1,35	1,49	1,28	1,57	1,21	1,65
35	1,40	1,52	1,34	1,58	1,28	1,65
40	1,44	1,54	1,39	1,60	1,34	1,66
45	1,48	1,57	1,43	1,62	1,28	1,67
50	1,50	1,59	1,46	1,63	1,42	1,67

Недостатками критерия Дарбина—Уотсона является наличие области неопределенности критерия, а также то, что критические значения *d*-статистики определены для объемов выборки не менее 15

▷ **Пример 14.4.** Выявить на уровне значимости 0,05 наличие автокорреляции возмущений для временного ряда y_t по данным табл. 14.1.

Р е ш е н и е. В примере **14.2** получено уравнение тренда $\tilde{y}_t = 181{,}32 + 25{,}679t$ (ед.). В табл. 14.3 приведен расчет сумм, необходимых для вычисления *d*-статистики.

Таблица 14.3

t	y_t	$\tilde{y}_t$	$e_t=y_t-\tilde{y}_t$	e_{t-1}	$e_t e_{t-1}$	e_t^2
1	213	207,0	6,0	—	—	36,0
2	171	232,7	−61,7	6,0	−370,2	3806,9
3	291	258,4	32,6	−61,7	−2011,4	1062,8
4	309	284,0	25,0	32,6	815,0	625,0
5	317	309,7	7,3	25,0	182,5	53,3
6	362	335,4	26,6	7,3	194,2	707,6
7	351	361,1	−10,1	26,6	−268,7	102,0
8	361	386,8	−25,8	−10,1	260,6	665,6
$\sum_{t=1}^{8}$	–	–	–	–	−1198,0	7059,2

Теперь по формуле (**14.10**) статистика $d \approx 2(1+1198{,}0/7059{,}2)=2{,}34$. По табл. 14.2 при $p=1$, $n=15$ критические значения $d_н=1{,}08$; $d_в=1{,}36$, т.е. фактически найденное $d=2{,}34$ находится в пределах от $d_в$ до $4-d_в$ $(1{,}36<d<2{,}64)$. Как уже отмечено, при $n<15$ критических значений *d*-статистики в табл. 14.2 нет, но судя по тенденции их изменений с уменьшением *n* можно предполагать, что найденное значение останется в интервале $(d_в;\ 4-d_в)$, т.е. для рассматриваемого временного ряда спроса на уровне значимости 0,05 гипотеза об отсутствии автокорреляции возмущений не отвергается (принимается). ▶

В случае отсутствия значимой автокорреляции возмущений методами регрессионного анализа может быть найдена не только точечная, но и интервальная оценка уровней ряда, т.е. осуществлены их *точечный* и *интервальный прогнозы*.

▷ **Пример 14.5.** По данным табл. 14.1 дать точечную и с надежностью 0,95 интервальную оценки прогноза среднего и индивидуального значений спроса на некоторый товар на момент $t = 9$ (девятый год).

Р е ш е н и е. Выше, в примере **14.2**, получено уравнение регрессии $\tilde{y}_t = 181{,}32+25{,}679t$, т.е. ежегодно спрос на товар увеличивался в среднем на 25,7 ед. Надо оценить условное математическое ожидание $M_{t=9}(Y)=\bar{y}(9)$. Оценкой $\bar{y}(9)$ является групповая средняя $\tilde{y}_{t=9} = 181{,}32+25{,}679\cdot 9=412{,}4$ ед.

Найдем по формуле **(13.6)** оценку s^2 дисперсии σ^2 (см табл. 14.3):

$$s^2=\frac{\sum_{t=1}^{8} e_t^2}{n-2}=\frac{7059,2}{8-2}=1176,5.$$

Вычислим оценку дисперсии групповой средней по формуле **(13.12)**:

$$s^2_{\tilde{y}_{t=9}}=1176,5\left(\frac{1}{8}+\frac{(9-4,5)^2}{42}\right)=714,3;\quad s_{\tilde{y}_{t=9}}=\sqrt{714,3}=26,73\,(\text{ед.})$$

(здесь мы использовали данные, полученные в примере **14.2**:

$$\bar{t}=\frac{\sum_{t=1}^{n} t}{n}=\frac{36}{8}=4,5;\ \sum_{t=1}^{n}(t-\bar{t})^2=\sum_{t=1}^{n} t^2-\frac{\left(\sum_{t=1}^{8} t\right)^2}{n}=$$

$$=204-\frac{36^2}{8}=42).$$

По табл. IV приложений $t_{0,95;6}$=2,45. Теперь по формуле **(13.13)** интервальная оценка прогноза среднего значения спроса:

$$412,4-2,45\cdot 26,73 \le \bar{y}(9)\le 412,4+2,45\cdot 26,73$$

или

$$346,9\le \bar{y}(9)\le 477,9\ (\text{ед.}).$$

Для нахождения интервальной оценки прогноза индивидуального значения $y^*(9)$ вычислим дисперсию его оценки по формуле **(13.14)**:

$$s^2_{y_{t_0=9}}=1176,5\left(1+\frac{1}{8}+\frac{(9-4,5)^2}{42}\right)=1890,8;\ s^2_{y_{t_0=9}}=43,48\ (\text{ед.}),$$

а затем по формуле **(13.15)** — саму интервальную оценку для $y^*(9)$:

$$412,4-2,45\cdot 43,48 \le y^*(9) \le 412,4+2,45\cdot 43,48$$

или

$$305,9 \le y^*(9) \le 518,9\ (\text{ед.}).$$

Итак, с надежностью 0,95 среднее значение спроса на товар на девятый год будет находиться в пределах от 346,9 до 477,9 (ед.), а его индивидуальное значение — от 305,9 до 518,9 (ед.). ▶

Прогноз развития изучаемого процесса на основе экстраполяции временных рядов может оказаться эффективным, как правило, в рамках *краткосрочного*, в крайнем случае, *среднесрочного периода прогнозирования.*

Если в рассматриваемой регрессионной модели автокорреляция возмущений существует, то необходимы меры по ее устранению (или снижению). С этой целью используются различные методы.

Метод последовательных разностей состоит, в частности, в переходе от уровней ряда y_t, x_t к их *первым* разностям

$$\Delta y_t = y_{t+1} - y_t,\ \Delta x_t = x_{t+1} - x_t\ (t = 1, 2,..., n) \qquad \textbf{(14.11)}$$

и рассмотрению уравнения регрессии $\Delta y_t = b_0 + b_1\Delta x_t$, в котором коэффициент b_1 интерпретируется как средний прирост переменной y_t при изменении прироста x_t на одну единицу. Метод эффективен, когда неслучайная составляющая временного ряда представляет прямую линию.

Другим возможным методом снижения автокорреляции является *включение в модель регрессии времени t* в качестве дополнительной объясняющей переменной:

$$y_t = b_0 + b_1x_t + b_2t. \qquad \textbf{(14.12)}$$

Метод оправдан, если он не приводит к мультиколлинеарности (см. § 13.9).

▷ **Пример 14.6.** По данным табл. 14.1, отражающей динамику цен x_t и спроса y_t некоторого товара за восьмилетний период, выяснить на уровне значимости 0,05, оказывает ли цена влияние на спрос.

Р е ш е н и е. Если абстрагироваться от того, что переменные x_t и y_t представляют собой временные ряды, то в предположении существования линейной регрессии можно получить аналогично примеру **12.1** (или **12.2**) уравнение регрессии в виде: $\tilde{y}_t = 635,2 - 0,8843x_t$. По F-критерию уравнение регрессии значимо на уровне 0,05, так как вычисленное значение статистики F=9,00>$F_{0,05;1;6}$=5,99.

Однако такой вывод был бы правомерен по крайней мере при отсутствии автокорреляции возмущений ε_t временного ряда зависимой переменной y_t. Использование критерия Дарбина—Уотсона показывает наличие существенной автокорреляции остаточного временного ряда $e_t=y_t-\tilde{y}_t$. Таким образом, для обоснованного ответа на вопрос задачи необходимо исключить автокорреляцию возмущений.

Первый способ. Перейдем в соответствии с **(14.11)** от уровней ряда y_t и x_t к их первым разностям Δy_t и Δx_t. Значения переменных Δx_t и Δy_t представим в табл. 14.4.

Таблица 14.4

Δx_t	−30	−112	−33	23	11	17	13
Δy_t	−42	120	18	8	45	−11	10

Рассчитанное уравнение регрессии: $\widetilde{\Delta y}_t = 9{,}94 - 0{,}7064\ \Delta x_t$.

Проверка уравнения по F-критерию на уровне 0,05 показывает, что оно незначимо, так как $F = 3{,}96 < F_{0{,}05;1;5} = 6{,}61$. Итак, нет оснований считать, что цена на данный товар оказывает существенное (значимое) влияние на спрос.

Второй способ. Включим в модель регрессии время t в качестве дополнительной объясняющей переменной. Методом наименьших квадратов (см. § 12.5) получим уравнение **(14.12)** в виде $y_t = 380{,}4 - 0{,}4443\ x_t + 19{,}22t$, причем коэффициент при переменной t оказался значимым по t-критерию ($t_{b_2} = 3{,}82 > t_{0{,}95;5} = 2{,}57$), а при переменной x_t — незначимым ($t_{b_1} = 2{,}22 < t_{0{,}95;5} = 2{,}57$). Следовательно, подтверждается вывод о незначимом влиянии цены x_t на спрос y_t на данный товар. ▶

Одним из распространенных методов устранения автокорреляции является использование авторегрессионной модели.

14.5. Авторегрессионная модель

Для данного временного ряда далеко не всегда удается подобрать *адекватную* модель, для которой ряд возмущений ε_t будет удовлетворять основным предпосылкам регрессионного анализа, в частности, не будет автокоррелирован.

До сих пор мы рассматривали модели вида **(14.5)**, в которых в качестве регрессора выступала переменная t — «время». В настоящее время достаточно широкое распространение получили и другие регрессионные модели, в которых регрессорами выступают *лаговые переменные*, т.е. переменные, влияние которых в регрессионной модели характеризуется некоторым запаздыванием. Еще одним отличием рассматриваемых в этом параграфе регрессионных моделей является то, что представленные в них объясняющие переменные являются величинами *случайными*.

Авторегрессионная модель p-го порядка имеет вид:

$$x_t = b_0 + b_1 x_{t-1} + b_2 x_{t-2} + \ldots + b_p x_{t-p} + \varepsilon_t \ (t = 1,2,\ldots,n), \qquad \textbf{(14.13)}$$

где b_0, $b_1,\ldots,b_p$ — некоторые константы.

Она описывает изучаемый процесс в момент t в зависимости от его значений в предыдущие моменты $t-1$, $t-2,\ldots$, $t-p$.

Если исследуемый процесс x_t в момент t определяется лишь его значениями в предшествующий период $t-1$, то рассматривают *авторегрессионную модель 1-го порядка* (*марковский случайный процесс*):

$$x_t = b_0 + b_1 x_{t-1} + \varepsilon_t \ (t = 1,2,\ldots,n). \qquad \textbf{(14.14)}$$

▷ **Пример 14.7.** В таблице представлены данные, отражающие динамику курса акций некоторой компании (ден. ед.):

Таблица 14.5

t	1	2	3	4	5	6	7	8	9	10	11
y_t	971	1166	1044	907	957	727	752	1019	972	815	823
t	12	13	14	15	16	17	18	19	20	21	22
y_t	1112	1386	1428	1364	1241	1145	1351	1325	1226	1189	1213

Используя авторегрессионную модель 1-го порядка, дать точечный и интервальный прогнозы среднего и индивидуального значений курса акций в момент $t = 23$, т.е. на глубину один интервал.

Р е ш е н и е. Попытка подобрать к данному временному ряду адекватную модель вида **(14.5)** с линейным или полиномиальным трендом оказывается бесполезной.

В соответствии с условием применим авторегрессионную модель вида **(14.14)**. Получим (аналогично примеру **14.2**)

$$\widetilde{y}_t = 284{,}0 + 0{,}7503 y_{t-1}. \qquad \textbf{(14.15)}$$

Найденное уравнение регрессии значимо на 5%-ном уровне по F-критерию, так как фактически наблюдаемое значение статистики $F = 24{,}32 > F_{0{,}05;1;19} = 4{,}35$. Применение критерия Дарбина—Уотсона свидетельствует о незначимой автокорреляции возмущений $e_t = y_t - \widetilde{y}_t$ (рекомендуем читателю убедиться в этом самостоятельно).

Вычисления, аналогичные примеру **14.5**, дают точечный прогноз по уравнению **(14.15)**: $\widetilde{y}_{t=23} = 284{,}0 + 0{,}7503 \cdot 1213 = 1194{,}1$ и интервальный на уровне значимости 0,05 для среднего и индивидуального значений — $1046{,}6 \le \overline{y}_{t=23} \le 1341{,}6$; $879{,}1 \le y^*_{t_0=23} \le 1509{,}1$.

Итак, с надежностью 0,95 среднее значение курса акций данной компании на момент $t = 23$ будет заключено в пределах от 1046,6 до 1341,6 (ден. ед.), а его индивидуальное значение — от 879,1 до 1509,1 (ден. ед.). ▶

В данной главе отражены лишь некоторые вопросы (элементы) анализа временных рядов. С более подробным их изложением можно ознакомиться, например, по [3], [24], а с анализом временных рядов на компьютере — [30].

Упражнения

В **примерах 14.8—14.10** имеются следующие данные об урожайности озимой пшеницы y_t (ц/га) за 10 лет:

t	1	2	3	4	5	6	7	8	9	10
y_t	16,3	20,2	17,1	7,7	15,3	16,3	19,9	14,4	18,7	20,7

14.8. Найти среднее значение, среднее квадратическое отклонение и коэффициенты автокорреляции (для лагов τ=1;2) временного ряда.

14.9. Найти уравнение тренда временного ряда y_t, полагая, что он линейный, и проверить его значимость на уровне 0,05.

14.10. Провести сглаживание временного ряда y_t методом скользящих средних, используя простую среднюю арифметическую с интервалом сглаживания: а) m=3; б) m=5.

14.11. В таблице представлены данные, отражающие динамику роста доходов на душу населения y_t (ден. ед.) за восьмилетний период:

t	1	2	3	4	5	6	7	8
y_t	1133	1222	1354	1389	1342	1377	1491	1684

Полагая тренд линейным: а) найти уравнение тренда и оценить его значимость на уровне 0,05; б) установить с помощью критерия Дарбина—Уотсона, является ли остаточный ряд e_t автокоррелированным на 5%-ном уровне значимости; в) при отсутствии автокорреляции возмущений дать точечный и с надежностью 0,95 интервальный прогнозы среднего и индивидуального значений доходов на девятый год.

Глава 15 Линейные регрессионные модели финансового рынка

Каждая ценная бумага — акция, облигация, контракт и другие — в каждый момент времени обладает стоимостью, которая называется *курсом* и устанавливается рынком (обычно как результат биржевых котировок). Даже обладая всей полнотой информации о выпустившем бумагу эмитенте, однозначно определить ее курс в заданный момент времени в будущем, как правило, невозможно. В этом случае наиболее естественно рассматривать курс ценной бумаги как значение случайной величины X.

Пусть X_t — курс ценной бумаги в момент времени t, а X_{t+1} — в момент времени t+1 (обычно единица времени — это промежуток между котировками). Обратимся к случаю, когда момент времени t уже наступил, а момент t+1 еще нет. Рассмотрим величину

$$r_t = \frac{X_{t+1} - X_t}{X_t}.$$

Так как X_{t+1} — случайная величина, то и r — тоже случайная величина. Она называется *доходностью ценной бумаги.*

Очевидно, что значение именно этой величины определяет привлекательность ценной бумаги для инвестора. И одна из главных задач финансового анализа состоит в возможно более точном предсказании значения величины r.

15.1. Регрессионные модели

Модели, рассматриваемые в финансовом анализе, связывают случайную величину r с величинами, которые объективно характеризуют финансовый рынок в целом. Такие величины называются *факторами.* В зависимости от постановки задачи факторы могут считаться как случайными, так и детерминированными, т.е. точно известными величинами.

В самом простом случае выделяется один фактор. Тогда статистическая модель имеет вид:

$$r = \alpha + \beta F + \varepsilon. \qquad \textbf{(15.1)}$$

Здесь α и β — постоянные (неизвестные параметры), ε — случайная величина, удовлетворяющая условию: $M_F(\varepsilon) = 0$, где

$M_F(\varepsilon)$ — условное математическое ожидание случайной величины ε относительно F. Из этого предположения следует, что и безусловное математическое ожидание величины ε также равно нулю. В самом деле:

$$M(\varepsilon) = M\big(M_F(\varepsilon)\big) = 0.$$

Отсюда также следует, что если фактор F рассматривается как случайная величина, то ее ковариация с ε равна нулю. Действительно, используя свойства условного математического ожидания, получаем:

$$\operatorname{cov}(F,\varepsilon) = M(F\varepsilon) - M(F)M(\varepsilon) = M(F\varepsilon) =$$
$$= M\big(M_F(F\varepsilon)\big) = M\big(FM_F(\varepsilon)\big) = 0.$$

Значения коэффициентов α и β нетрудно выразить через числовые характеристики r и F:

$$\operatorname{cov}(r,F) = \beta\operatorname{cov}(F,F) + \operatorname{cov}(\varepsilon,F)$$

или

$$\operatorname{cov}(r,F) = \beta\operatorname{cov}(F,F).$$

Отсюда

$$\beta = \frac{\operatorname{cov}(r,F)}{\operatorname{cov}(F,F)} = \frac{\operatorname{cov}(r,F)}{D(F)}.$$

Перейдя в уравнении модели **(15.1)** к математическим ожиданиям, получим:

$$M(r) = \alpha + \beta M(F) + M(\varepsilon).$$

Но $M(\varepsilon)=0$, поэтому

$$\alpha = M(r) - \beta M(F) = M(r) - \frac{\operatorname{cov}(r,F)}{D(F)}M(F).$$

Коэффициент β называется *чувствительностью доходности* ценной бумаги к фактору F. Коэффициент α называется *сдвигом*.

В классическом регрессионном анализе значения факторов F считаются детерминированными величинами, т.е. модель **(15.1)** имеет вид:

$$r_t = \alpha + \beta F_t + \varepsilon_t.$$

Здесь $t = 1,\ldots,n$ — моменты времени — интерпретируются как номер наблюдения; $F_1,\ldots, F_n$ — известные значения факторов; r_t — наблюдаемые выборочные значения случайной величины r; α и β — неизвестные параметры. Их оценки[1] можно построить методом наименьших квадратов (см. гл. 12):

$$\hat{\beta} = \frac{\hat{\operatorname{cov}}(r,F)}{\hat{D}(F)} = \frac{n\sum_{t=1}^{n} F_t r_t - \left(\sum_{t=1}^{n} F_t\right)\left(\sum_{t=1}^{n} r_t\right)}{n\sum_{t=1}^{n} F_t^2 - \left(\sum_{t=1}^{n} F_t\right)^2},$$

$$\hat{\alpha} = \hat{M}(r) - \frac{\hat{\operatorname{cov}}(r,F)}{\hat{D}(F)}\hat{M}(F) = \frac{1}{n}\sum_{t=1}^{n} r_t - \frac{1}{n}\left(\sum_{t=1}^{n} F_t\right)\hat{\beta}.$$

Разные модели финансового рынка рассматривают различные величины в качестве фактора F. Рассмотрим далее основные из этих моделей.

15.2. Рыночная модель

Одна из самых распространенных моделей использует в качестве фактора F доходность рыночного индекса r_M. *Рыночным индексом* называется взвешенная сумма курсов акций наиболее значительных эмитентов финансового рынка. Например, в США наиболее распространены следующие индексы:

DJ (*индекс Доу—Джонса*) — рассчитывается по 30 наиболее значимым корпорациям, таким как Microsoft, Coca Cola, General Motors и т.д.;

индекс S&P 500 (*Standard and Poor's*) — рассчитывается по 500 наиболее крупным компаниям;

сводный индекс NYSE — для его расчета используются курсы акций, зарегистрированных на Нью-Йоркской фондовой бирже.

Очевидно, рыночный индекс в определенной степени отражает состояние экономики в целом. Так что рыночная модель показывает, насколько доходность ценной бумаги соответствует экономической динамике страны (или даже сообщества стран).

Случайная величина ε отражает зависимость доходности ценной бумаги от обстоятельств, специфических именно для ее эмитента.

Обсудим смысл коэффициента β в случае рыночной модели.

Доходность рыночного индекса по сути представляет собой усредненную доходность различных ценных бумаг. Если рассматривать множество всех ценных бумаг, фигурирующих на рынке, то коэффициент β наудачу выбранной ценной бумаги представ-

[1] Оценки параметров обозначаем со знаком ^.

ляет собой значение случайной величины (обозначим ее той же буквой β). Если ценная бумага включена в индекс, то по самому определению индекса имеет место равенство: $M(\beta) = 1$.

Если конкретное наблюдаемое значение ценной бумаги больше единицы, значит, ее доходность растет в среднем быстрее, чем рынок в целом. Такие бумаги называются *«агрессивными»*; бумаги с коэффициентом β, меньшим единицы, называются *«оборонительными»*.

▷ **Пример 15.1.** Значения доходности акций Widget Manufacturing и доходности индекса даны в табл. 15.1:

Таблица 15.1 [34]

Год	*Квартал*	*Доходность WM*	*Доходность индекса*
1	1	−13,38	2,52
	2	16,79	5,45
	3	−1,67	0,76
	4	−3,46	2,36
2	5	10,22	8,56
	6	7,13	8,67
	7	6,71	10,80
	8	7,84	3,33
3	9	2,15	−5,07
	10	7,95	7,10
	11	−8,05	−11,57
	12	7,68	4,65
4	13	4,75	14,59
	14	7,55	2,66
	15	−2,36	3,81
	16	4,98	7,99

Оценить зависимость доходности акций от доходности индекса.

Р е ш е н и е. Метод наименьших квадратов дает следующие результаты: $\hat{\beta}=0{,}63$, $\hat{\alpha}=0{,}79$. Стандартные ошибки $\sigma_{\hat{\beta}}=0{,}28$, $\sigma_{\hat{\alpha}}=2{,}03$, коэффициент детерминации $R^2=0{,}27$. Полученные характеристики свидетельствуют о том, что доходность акций Widget Manufacturing существенно зависит от рыночного индекса. Как видно, акции Widget Manufacturing являются «оборонительными». ▶

На финансовом рынке присутствуют и бумаги с нулевым коэффициентом β. В этом случае имеет место соотношение $r = r_f$, откуда следует, что ожидаемая доходность ценной бумаги фиксированная и не зависит от состояния рынка в целом. Такая ситуация характерна, например, для облигаций.

15.3. Модели зависимости от касательного портфеля

Другим фактором, часто используемым в линейных регрессионных моделях, является доходность некоторого выделенного портфеля ценных бумаг, который называется *касательным*. Это понятие было введено Г. Марковицем в 1952 г. Опишем это понятие.

Пусть на финансовом рынке обращается n ценных бумаг и капитал, равный единице, инвестируется в эти бумаги так, что x_i — капитал, инвестируемый в i-ю бумагу. Набор чисел $p = x_1,...,x_n$, удовлетворяющий условию $x_1 + x_2 + ... + x_n = 1$, назовем *портфелем ценных бумаг*. Разумеется, некоторые числа $\{x_i\}$ могут быть нулевыми. (На самом деле некоторые x_i могут быть и отрицательными: соответствующая ситуация называется *продажей ценной бумаги без покрытия*. Мы, однако, не будем углубляться в это понятие).

Каждому портфелю p соответствует случайная величина r_p — доходность, которая определяется аналогично доходности одной ценной бумаги. Очевидно, $r_p = \sum_{i=1}^{n} x_i r_i$. Рассмотрим координатную плоскость, на которой по оси ординат откладывается математическое ожидание доходности (ожидаемая доходность $\bar{r}_p$), а по оси абсцисс — стандартное отклонение доходности — $\sigma_p = \sqrt{D(r_p)}$. Величина σ_p называется *риском портфеля*. Тогда каждому портфелю может быть поставлена в соответствие точка на такой координатной плоскости, а все множество допустимых портфелей отображается в некоторую двумерную фигуру, называемую *допустимым множеством* (рис. 15.1).

Между множеством всех портфелей и допустимым множеством, разумеется, нет взаимно-однозначного соответствия. Конечно, два различных портфеля могут иметь равные значения $\bar{r}_p$ и σ_p.

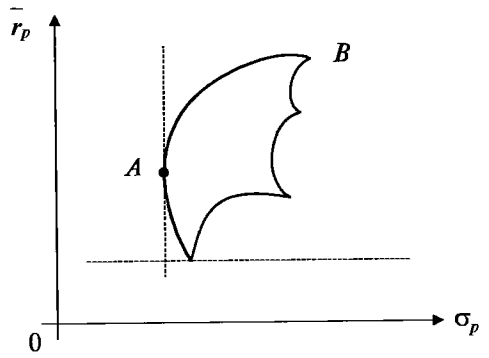


Рис. 15.1

Естественно предположить, что инвестор предпочитает получить бóльшую доходность с наименьшим риском, т.е. из двух портфелей с одинаковым значением $\bar{r}_p$ он выберет тот, значение σ_p которого меньше. Это значит, что наиболее предпочтительному портфелю соответствует точка на куске границы AB (см. рис. 15.1). Линия AB называется *эффективным множеством.*

Проблема выбора точки эффективного множества решается каждым инвестором индивидуально и, казалось бы, зависит от его склонности к риску (или, наоборот — к избеганию риска). Оказывается, однако, что эффективному множеству принадлежит точка, которая является выделенной для всех инвесторов.

Предположим, что кроме приобретения ценных бумаг инвестор имеет возможность безрискового предоставления и получения займов. Такое предположение вполне соответствует действительности, если инвестор имеет возможность покупать государственные облигации и брать кредит. Мы сделаем еще одно предположение (уже отнюдь не столь бесспорное), что безрисковое предоставление и получение займов происходит с одной и той же процентной ставкой r_f, которая называется *безрисковой ставкой.*

Рассмотрим прямую l, пересекающую ось ординат в точке r_f и касательную к эффективному множеству.

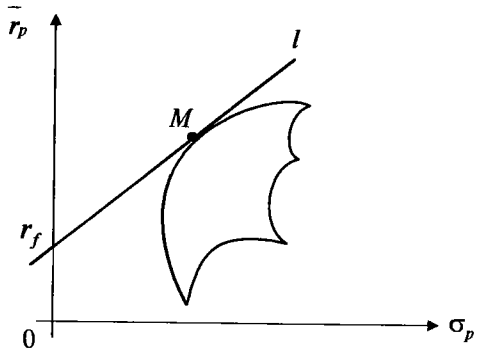


Рис. 15.2

Уравнение прямой l имеет вид $\bar{r} = r_f + \frac{\sigma}{\sigma_M}(\bar{r}_M - r_f)$.

Рассмотрим портфель $p = (x_f, x_M)$, где x_f — безрисковые вложения (положительные в случае приобретения облигаций и отрицательные при заеме средств) с фиксированной ставкой r_f; x_M — вложение в портфель, соответствующий точке M. Тогда $\bar{r}_p = x_f r_f + x_M \bar{r}_M = (1 - x_M) r_f + x_M \bar{r}_M = r_f + x_M(\bar{r}_M - r_f)$; $\sigma_p = x_M \sigma_M$, откуда $\bar{r}_p = r_f + \frac{\sigma_p}{\sigma_M}(\bar{r}_M - r_f)$, т.е. точка $(\sigma_p, \bar{r}_p)$ лежит на прямой l. Очевидно также, что любая точка полупрямой l, лежащая в первой четверти, достижима с помощью комбинации (x_f, x_M).

Таким образом, при наличии возможности безрискового предоставления и получения займов допустимое множество расширяется, а эффективным множеством становится прямая l.

Портфель, соответствующий точке касания M (рис.15.2), называется *касательным портфелем.*

Таким образом, оптимальной для любого инвестора стратегией оказывается инвестирование части средств в касательный портфель, а части — в безрисковые облигации. Либо наоборот: получение займа для дополнительного инвестирования в касательный портфель.

Разумеется, на практике точное нахождение касательного портфеля невозможно. Но для многих практических целей оказывается полезной модель, в которой в качестве фактора выбрана доходность касательного портфеля, а точнее — разница между r_M и безрисковой ставкой r_f. Таким образом, $F = r_M - r_f$ и модель имеет вид:

$$r_i = \alpha_i + \beta_i(r_M - r_f) + \varepsilon_i,$$

где i — номер ценной бумаги.

15.4. Неравновесные и равновесные модели

Очевидно, что доходности ценных бумаг, обращающихся на рынке, можно рассматривать в зависимости от времени. При этом будут зависеть от времени числовые характеристики случайной величины r_p. Так же, вообще говоря, будут зависеть от времени и значения параметров α и β.

Модель финансового рынка называется *равновесной*, если числовые характеристики входящих в нее случайных величин постоянны во времени. Экономический смысл подобного предположения очевиден: рынок считается «устоявшимся», сбалансированным. В этом случае можно получить некоторые конкретные результаты, существенно упрощающие ситуацию.

Будем рассматривать модель зависимости доходности ценной бумаги от доходности касательного портфеля (предполагается, что безрисковая ставка получения и предоставления займов для всех участников рынка одна и та же и равна r_f). *Если модель равновесная, т.е. рынок сбалансированный, то касательный портфель удовлетворяет следующему свойству: доля каждой ценной бумаги в нем соответствует ее относительной рыночной стоимости.* Такой портфель называется *рыночным* и определяется однозначно. Таким образом, рассматривая равновесные модели, мы будем отождествлять понятия касательного и рыночного портфеля, доходность которого обозначим r_M.

Итак, регрессионная модель для i-й ценной бумаги имеет вид:

$$r_i = \alpha_i + \beta_{iM}(r_M - r_f) + \varepsilon_i.$$

Оказывается, в равновесном случае имеет место следующая теорема.

Теорема. *Для всех ценных бумаг, обращающихся на рынке, коэффициент α_i один и тот же и равен безрисковой ставке.*

□ Имеем

$$\bar{r}_i = \alpha_i + \beta_{iM}(\bar{r}_M - r_f).$$

Рассмотрим портфель p, состоящий из i-й ценной бумаги и рыночного портфеля M в пропорции x_i и $1-x_i$ соответственно. Ожидаемая доходность такого портфеля составит

$$\bar{r}_p = x_i\bar{r}_i + (1-x_i)\bar{r}_M, \qquad (15.2)$$

а стандартное отклонение будет

$$\sigma_p = \sqrt{x_i^2\sigma_i^2 + (1-x_i)^2\sigma_M^2 + 2x_i(1-x_i)\sigma_{iM}}. \qquad (15.3)$$

Все такие портфели отображаются на кривую, соединяющую точки i и M (рис. 15.3).

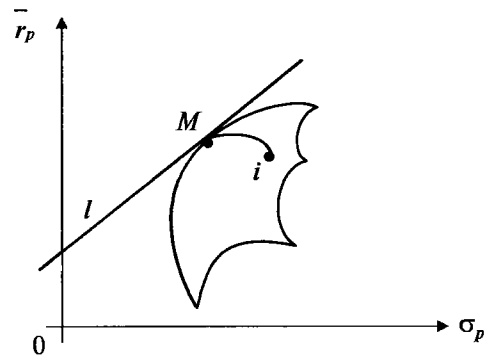


Рис. 15.3

Из равенства **(15.2)** получаем:

$$\frac{d\bar{r}_p}{dx_i} = \bar{r}_i - \bar{r}_M.$$

А из равенства **(15.3)**:

$$\frac{d\sigma_p}{dx_i} = \frac{x_i\sigma_i^2 - \sigma_M^2 + x_i\sigma_M^2 + \sigma_{iM} - 2x_i\sigma_{iM}}{\sqrt{x_i^2\sigma_i^2 + (1-x_i)^2\sigma_M^2 + 2x_i(1-x_i)\sigma_{iM}}},$$

откуда

$$\frac{d\bar{r}_p}{d\sigma_p} = \frac{(\bar{r}_i - \bar{r}_M)\sqrt{x_i^2\sigma_i^2 + (1-x_i)^2\sigma_M^2 + 2x_i(1-x_i)\sigma_{iM}}}{x_i\sigma_i^2 - \sigma_M^2 + \sigma_{iM} - 2x_i\sigma_{iM}}.$$

В точке M $x_i = 0$, отсюда наклон кривой в точке M:

$$\frac{d\bar{r}_p}{d\sigma_p} = \frac{(\bar{r}_i - \bar{r}_M)\sigma_M}{\sigma_{iM} - \sigma_M^2}. \qquad (15.4)$$

Но кривая касается прямой l, поэтому

$$\frac{d\bar{r}_p}{d\sigma_p} = \frac{\bar{r}_M - r_f}{\sigma_M}. \qquad (15.5)$$

Приравняв правые части равенств **(15.4)** и **(15.5)**, получим

$$\bar{r}_i = r_f + \left[\frac{\bar{r}_M - r_f}{\sigma_M^2}\right]\sigma_{iM}. \blacksquare$$

Таким образом, *единственным параметром, характеризующим данную ценную бумагу, является ее чувствительность «бета» к рыночному портфелю.*

15.5. Модель оценки финансовых активов (CAPM)[1]

Уравнение

$$\bar{r}_i = r_f + (\bar{r}_M - r_f)\beta_{iM} \qquad (15.6)$$

называется *рыночной линией* ценной бумаги. Оно определяет зависимость ожидаемой доходности ценной бумаги от ее чувствительности «бета» $\left(\beta_{iM} = \sigma_{iM}/\sigma_M^2\right)$.

Рассмотрим портфель

$$\{x_1,\ldots,x_n\},\ \sum_{i=1}^{n} x_i = 1.$$

Доходность портфеля

$$r_p = \sum_{i=1}^{n} x_i r_i.$$

Отсюда имеем, что ожидаемая доходность

$$\bar{r}_p = \sum_{i=1}^{n} x_i \bar{r}_i = \sum_{i=1}^{n} x_i r_f + (\bar{r}_M - r_f)\sum_{i=1}^{n} x_i \beta_{iM} = r_f + (\bar{r}_M - r_p)\beta_{pM}.$$

[1] CAPM — Capital Asset Prising Model.

Здесь

$$\beta_{pM} = \frac{\operatorname{cov}(r_p, r_M)}{\sigma_M^2} = \frac{\sum_{i=1}^{n} x_i \operatorname{cov}(r_i, r_M)}{\sigma_M^2}.$$

Уравнение

$$\bar{r}_p = r_f + (\bar{r}_M - r_p)\beta_{pM}$$

называется *уравнением модели оценки финансовых активов*. Для ее использования необходимо получить оценки параметров касательного портфеля — ожидаемой доходности и риска, а также ковариаций доходностей ценных бумаг, входящих в p, с доходностью рыночного портфеля.

Практическое значение модели оценки финансовых активов заключается в том, что она может служить для выявления неверно оцененных бумаг в неравновесной ситуации, т.е. в ситуации несбалансированного рынка. Так, если доходность ценной бумаги выше той, которая задается уравнением **(15.6)**, то бумага является *переоцененной*, в противоположном случае — *недооцененной*.

15.6. Связь между ожидаемой доходностью и риском оптимального портфеля

Пусть для всех участников рынка имеется единая безрисковая ставка r_f получения и предоставления займов. Тогда *оптимальным* оказывается *портфель, представляющий собой комбинацию касательного и безрискового*. Оптимальный портфель имеет вид:

$$p_{\text{опт}} = (x_0, x_1, x_2,\ldots, x_n).$$

Здесь x_0 — часть (доля) средств, вложенных в безрисковые бумаги (казначейские векселя), если $x_0>0$, или x_0 — заемные средства, использованные для вложения в касательный портфель, если $x_0<0$. Числа $x_1, x_2,\ldots,x_n$ находятся в том же соотношении, в котором находятся компоненты касательного портфеля.

Очевидно, что доходности оптимального и касательного портфелей связаны линейным соотношением:

$$r_p = x_0 r_f + \lambda r_M,$$

откуда следует, что коэффициент корреляции $\rho(r_p, r_M) = 1$, т.е.

$$\frac{\operatorname{cov}(r_p, r_M)}{\sigma_p \sigma_M} = 1 \quad \text{и} \quad \beta_{pM} = \frac{\sigma_p}{\sigma_M}.$$

Таким образом, уравнение CAPM имеет вид:

$$\bar{r}_p = r_f + \frac{(\bar{r}_M - r_f)}{\sigma_M}\sigma_p. \qquad (15.7)$$

Оно задает зависимость ожидаемой доходности оптимального портфеля от его риска.

Рассмотрим геометрический смысл уравнения (15.7). При сделанных предположениях о безрисковой ставке r_f эффективное множество представляет собой прямую l (рис. 15.4).

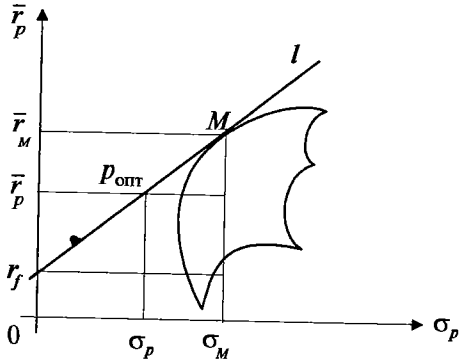


Рис. 15.4

Из очевидного подобия треугольников следует:

$$\frac{\bar{r}_p - r_f}{\sigma_p} = \frac{\bar{r}_M - r_f}{\sigma_M},$$

что эквивалентно равенству (15.7).

15.7. Многофакторные модели

Однофакторные модели во многих случаях являются вполне адекватными, однако чаще всего они оказываются слишком упрощенными и тогда приходится рассматривать зависимость доходности ценной бумаги от нескольких (m) факторов, т.е. линейные регрессионные модели вида:

$$r_i = \alpha + \sum_{k=1}^{m} \beta_k F_i^{(k)} + \varepsilon_i.$$

Здесь α, β_k — параметры, $F^{(k)}$ — факторы, определяющие состояние рынка (i — номер наблюдения).

Такими факторами могут быть, например, уровень инфляции, темпы прироста валового внутреннего продукта и др. Если данная ценная бумага относится к некоторому сектору экономики, то безусловно следует рассматривать факторы, специфические для данного сектора.

▷ **Пример 15.2.** В табл. 15.2 приведены данные за шесть лет о темпе роста, уровне инфляции и доходности акций компаний Widget Manufacturing.

Таблица 15.2 [34]

Год	*Темп роста ВВП, % (temp)*	*Уровень инфляции, % (inf)*	*Доходность акций, % (Widget)*
1-й	5,7	1,1	14,3
2-й	6,4	4,4	19,2
3-й	7,9	4,4	23,4
4-й	7,0	4,6	15,6
5-й	5,1	6,1	9,2
6-й	2,9	3,1	13,0

Рассматривается факторная модель вида:

$$\text{Widget} = \alpha + \beta_1 \text{temp} + \beta_2 \text{inf}.$$

С помощью метода наименьших квадратов получается уравнение вида:

$$\overline{\text{Widget}} = 5{,}77 + 2{,}17\,\text{temp} - 0{,}67\text{inf}.$$

При этом средние квадратические ошибки оценок параметров β_1 и β_2 равны 1,125 и 1,162.

Следует стремиться к возможно меньшему количеству объясняющих переменных (факторов), поскольку кроме усложнения модели «лишние» факторы приводят к увеличению ошибок оценок. Так, в рассмотренном примере стандартная ошибка оценки параметра β_2 оказалась больше ее значения по абсолютной величине. Это наводит на мысль, что четкой зависимости доходности акций компании Widget от инфляции нет. В этом случае естественно удалить фактор инфляции и рассматривать

зависимости доходности Widget только от темпа роста ВВП. Соответствующее уравнение регрессии

$$\overline{\text{Widget}} = \alpha + \beta_1 \text{ temp},$$

оцениваемое с помощью метода наименьших квадратов, имеет вид:

$$\overline{\text{Widget}} = 4+2 \text{ temp}.$$

При этом ошибка параметра β_1 уменьшается и становится равной 1,002.

Однако в реальных ситуациях порой приходится рассматривать модели зависимости от десятков и даже сотен факторов.

В 1970-х гг. в США специалисты по эконометрике Б. Розенберг и В. Марат сформулировали сложную многофакторную модель, связывавшую доходность акций с множеством факторов, полученных из данных по деловым операциям соответствующих компаний. В дальнейшем Розенберг основал фирму, которая теперь называется BARRA, с целью развития модели и ее продажи институциональным инвесторам. Основная модель BARRA учитывает 68 основных факторов. В настоящее время BARRA выросла во всемирную консалтинговую организацию с ежегодным доходом более 40 млн долл

Библиографический список

1. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Основы моделирования и первичная обработка данных. — М.: Финансы и статистика, 1983.
2. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Исследование зависимостей. — М.: Финансы и статистика, 1985.
3. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. — М.: ЮНИТИ, 1998.
4. *Бочаров П.П., Печенкин А.В.* Теория вероятностей и математическая статистика. — М.: Гардарика, 1998.
5. *Вентцель Е.С.* Исследование операций. Задачи, принципы, методология. — М.: Наука, 1990.
6. *Вентцель Е.С., Овчаров Л.А.* Задачи и упражнения по теории вероятностей. — М.: Высшая школа, 2002.
7. *Вентцель Е.С., Овчаров Л.А.* Теория вероятностей и ее инженерные приложения. — М.: Наука, 1988.
8. *Войтенко М.А.* Руководство к решению задач по теории вероятностей. — М.: Изд. ВЗФЭИ, 1988.
9. *Высшая математика* для экономистов/Под ред. Н.Ш. Кремера — М.: ЮНИТИ, Банки и биржи, 1998.
10. *Гмурман В.Е.* Руководство к решению задач по теории вероятностей и математической статистике. — М.: Высшая школа, 1979.
11. *Гмурман В.Е.* Теория вероятностей и математическая статистика. — М.: Высшая школа, 1977.
12. *Гнеденко Б.В.* Курс теории вероятностей. — М.: Наука, 1975.
13. *Дубров А.М., Мхитарян В.С., Трошин Л.И.* Многомерные статистические методы. — М.: Финансы и статистика, 1998.
14. *Иванова В.М., Калинина В.Н.* и др. Математическая статистика. — М.: Высшая школа, 1981.
15. *Исследование операций в экономике*/Под ред. Н.Ш. Кремера. — М.: ЮНИТИ, Банки и биржи, 1997.
16. *Калинина В.Н., Панкин В.Н.* Математическая статистика. — М.: Высшая школа, 1998.
17. *Карасев А.И., Аксютина З.М., Савельева Т.И.* Курс высшей математики для экономических вузов. — М.: Высшая школа, 1982, ч. 2.

18. *Карасев А.И., Калихман И.Л., Кремер Н.Ш.* Матричная алгебра. — М.: Изд. ВЗФЭИ, 1987.

19. *Карасев А.И., Кремер Н.Ш., Савельева Т.И.* Математические методы и модели в планировании. — М.: Экономика, 1987.

20. *Колемаев В.А., Калинина В.Н.* Теория вероятностей и математическая статистика. — М.: ИНФРА—М, 1997.

21. *Колмогоров А.Н.* Основные понятия теории вероятностей. — М.: Наука, 1975.

22. *Крамер Г.* Математические методы статистики: Пер. с англ. — М.: Мир, 1975.

23. *Кремер Н.Ш.* Математическая статистика. — М.: Экономическое образование, 1992.

24. *Кремер Н.Ш., Путко Б.А.* Эконометрика. — М.: ЮНИТИ-ДАНА, 2002.

25. *Математика* в экономике: учебно-методическое пособие /Под ред. Н.Ш. Кремера. — М.: Финстатинформ, 1999.

26. *Мацкевич И.П., Свирид Г.П., Булдык Г.М.* Сборник задач и упражнений по высшей математике. Теория вероятностей и математическая статистика. — Минск: Вышэйшая школа, 1996.

27. *Мацкевич И.П., Свирид Г.П.* Высшая математика. (Теория вероятностей и математическая статистика). — Минск, Вышэйшая школа, 1993.

28. *Смирнов Н.В., Дунин-Барковский И.В.* Курс теории вероятностей и математической статистики для технических приложений. — М.: Наука, 1969.

29. *Справочник* по прикладной статистике/Под ред. Э. Ллойда, У. Лидермана: Пер. с англ. — М.: Финансы и статистика, 1989.

30. *Тюрин Ю.Н., Макаров А.А.* Статистический анализ данных на компьютерах/Под ред. В.Э. Фигурнова. — М.: ИНФРА—М, 1998.

31. *Уотшем Т. Дж., Паррамоу К.* Количественные методы в финансах: Пер. с англ. — М.: ЮНИТИ, Финансы, 1999.

32. *Ферстер Э., Ренц Б.* Методы корреляционного и регрессионного анализа: Пер. с нем. — М.: Финансы и статистика, 1983.

33. *Хьютсон А.* Дисперсионный анализ: Пер. с англ. — М.: Статистика, 1971.

34. *Четыркин Е.М., Калихман И.Л.* Вероятность и статистика. — М.: Финансы и статистика, 1982.

35. *Шарп У., Гордон Дж. А., Бейли Д.* Инвестиции: Пер. с англ. — М.: ИНФРА—М, 1997.

36. *Экономико-математические* методы и прикладные модели/Под ред. В.В. Федосеева. — М.: ЮНИТИ, 1999.

Ответы к упражнениям

Глава 1

1.37. а) $1/P_7=1/7!=1/5040=0{,}000198$; б) $P_2P_3P_2P_2/P_{10} = 2!3!2!2!/10! = 1/75600 = 0{,}0000132$.

1.38. $1/P_5 = 1/5! = 1/120 = 0{,}00833$.

1.39. а) $C_{15}^4/C_{25}^4 = 0{,}108$; б) $C_{10}^4/C_{25}^4 = 0{,}0166$; в) $C_{15}^1 C_{10}^3/C_{25}^4 = 0{,}142$.

1.40. а) $C_{10}^3 C_{10}^2/C_{20}^5 = 0{,}348$; б) $1 - C_{10}^5/C_{20}^5 = 0{,}984$.

1.41. $C_3^1 C_2^1/C_5^2 = 0{,}6$.

1.42. $\left(C_{20}^3C_{10}^2 + C_{20}^4C_{10}^1 + C_{20}^5\right) / C_{30}^5 = 0{,}809$.

1.43. а) $1/(0{,}9 \cdot A_{10}^5) = 1/27216 = 0{,}0000367$; б) $9/90000 = 0{,}0001$; в) $5^5/90000 = 5/144 = 0{,}0347$.

1.44. а) $C_2^1 C_{14}^7/C_{16}^8 = 8/15 = 0{,}533$; б) $\left(C_{14}^8 + C_2^2C_{14}^6\right)/C_{16}^8 = 7/15 = 0{,}467$.

1.45. а) $\left(C_{19}^2C_5^1 + C_{19}^3\right)/C_{24}^3 = 0{,}901$; б) $0{,}099$.

1.46. $C_4^3C_2^2C_2^1/C_{10}^6 = 0{,}038$.

1.47. $P_8P_3/P_{10} = 8!3!/10! = 0{,}067$.

1.48. а) $0{,}125$; б) $0{,}5$.

1.49. $\left(1 - C_5^3/C_8^3\right)\left(1 - C_3^2/C_5^2\right) = 0{,}575$.

1.50. $0{,}4375$.

1.51. $2/\pi = 0{,}637$.

1.52. $C_{95}^{50} / C_{100}^{50} = 0{,}0281$.

1.53. $0{,}571$.

1.54. а) $0{,}032$; б) $0{,}316$.

1.55. $0{,}316$.

1.56. $0{,}788$.

1.57. а) $0{,}54$; б) $0{,}995$.

1.58. а) $0{,}46$; б) $0{,}7$.

1.59. $0{,}982$.

1.60. $C_4^1C_{16}^2 / C_{20}^3 = 0{,}421$.

1.61. $1-(C_8^0C_{12}^3 + C_8^1C_{12}^2)/C_{20}^3 = 0{,}344$.

1.62. $1-C_{11}^2 / C_{15}^2 = 0{,}476$.

1.63. а) $(C_7^3 + C_{13}^3)/C_{20}^3 = 0{,}282$;

б) $(C_7^2C_{13}^1 + C_7^3C_{13}^0)/C_{20}^3 = 0{,}270$.

1.64. а) $C_{10}^1C_3^1C_7^1 / C_{20}^3 = 0{,}184$;

б) $(C_{10}^3 + C_3^3 + C_7^3)/C_{20}^3 = 0{,}137$.

1.65. а) 0,0104; б) 0,625.
1.66. Не менее 5 пакетов.
1.67. $n \geq \lg(1-\mathcal{P}) / \lg(1-p)$; при p=0,4, $\mathcal{P}$=0,8704, $n \geq 4$.
1.68. $1-C_7^3 / C_{10}^3 = 0{,}708$.
1.69. 0,4.
1.70. 0,992.
1.71. 0,664.
1.72. а) 0,03; б) 0,167.
1.73. а) 0,1725; б) 0,317.
1.74. 0,667.
1.75. 0,621.
1.76. а) 0,9236; б) 0,0022.
1.77. 0,0073.
1.78. Вероятность одна и та же, равная m/n.
1.79. а) $1/6^3$=0,0046;
б) $6/6^3$=1/36 ≈ 0,028;
в) $C_6^3 / 6^3 = 5/54 \approx 0{,}093$.
1.80. а) $5/5^3$=0,04;
б) $C_5^3 \cdot 3!/5^3 = 0{,}48$.
1.81. а) 2/19≈0,105;
б) $(2 \cdot 18!+2 \cdot 18 \cdot 18!)/20!$= =0,1.
1.82. $1 \cdot (6/9)(5/8)(4/7)(3/9) \times \times (2/8)(1/7)$=5/1764≈ ≈0,0028.
1.83. а) $0{,}7(1-0{,}2 \cdot 0{,}15 \cdot 0{,}1)$= =0,6979;
б) $0{,}7 \cdot 0{,}2 \cdot 0{,}85$=0,119;
в) $0{,}7 \cdot 0{,}8 \cdot 0{,}85 \cdot 0{,}9$= =0,4284.
1.84. $1/6+(5/6) \cdot (3/51)$=11/51≈ ≈0,216.
1.85. $2/5+(3/5)(2/4)(2/3)$=0,6.
1.86. $3 \cdot 0{,}8^2 \cdot 0{,}2^2$=0,0768.
1.87. $(60/100)(59/99)+(60/100) \times \times (40/99)(59/58)$≈0,504.
1.88. $0{,}2 \cdot ,9/(1-0{,}8 \cdot 0{,}9)$= =9/14=0,643.
1.89. а) ≈0,579; б) ≈0,002.

Глава 2

2.13. а) $P_{2,6} = C_6^2 p^2 q^4 = 0{,}015$; б) $P_6(m>2) = 1-P_6(m\leq 2)$= $=1-(P_{0,6}+P_{1,6}+ P_{2,6}) = 0{,}999$.
2.14. а) $P_{3,10}=C_{10}^3 p^3 q^7 = 0{,}201$; б) $P_{10}(m<3) = P_{0,10} + P_{1,10} + P_{2,10}$=0,678.
2.15. $P_6(m \geq 4) = P_{4,6} + P_{5,6} + P_{6,6} = 0{,}544$.
2.16. а) $P_{3,10}= C_{10}^3 p^3 q^7 = 0{,}1298$; б) $P_{10}(m<2)=P_{0,10}+P_{1,10}$=0,544.
2.17. $P_6(m \leq 2) = P_{0,6}+P_{1,6}+P_{2,6}$=0,984.
2.18. а) $P_{10}(m \geq 3) = 1 - P_{10}(m<3) = 1 - (P_{0,10}+P_{1,10}+P_{2,10}) = 0{,}945$;
б) $P_{10}(m \leq 3) = P_{0,10} + P_{1,10} + P_{2,10} + P_{3,10} = 0{,}172$.
2.19. а) 2 партии из 4, ибо $P_{2,4} = C_4^2 p^2 q^2 = 0{,}375$, а $P_{3,6,} = C_6^3 p^3 q^3$ = =0,312; б) не менее 2 партий из 4, ибо $P_4(m \geq 2)$ = $= P_{2,4} + P_{3,4} + P_{4,4} = 0{,}688$, а $P_6(m \geq 3) = 1 - P_6(m < 3) = 1 -$ $-(P_{0,6} + P_{1,6} + P_{2,6}) = 0{,}656$.

2.20. а) $P_{3,400}=P_3(0,4) = 0{,}0072$; б) $P_{400}(m\leq 3) = P_{0,400} + P_{1,400} + P_{2,400} + +P_{3,400} = 0{,}9992$.
2.21. а) $P_{48,100000} = 0{,}054$; б) $P_{100000}(45 \leq m \leq 55) = P_{100000}$ $(|m - np| \leq 5) = = 0{,}522$.
2.22. а) $m_0 = 10$; $P_{10,3650}=P_{10}(10)=0{,}1251$; $P_{3650}(3\leq m\leq 3650)=0{,}9971$
2.23. а) $P_{5,10000} = P_5(1) = 0{,}0031$; б) $P_{10000}(m \geq 9998)$ при p = = 0,9999 равно $P_{10000}(m \leq 2)$ при p = 0,001, т.е. $P_{10000}(m \leq 2) = P_{0,10000} + P_{1,10000} + P_{2,10000} = P_0(1)+P_1(1)+ +P_2(1) = 0{,}9197$.
2.24. а) $q_1^3 q_2^2 + \left(C_3^1 p_1 q_1^2\right)\left(C_3^1 p_2 q_2^2\right) + \left(C_3^2 p_1^2 q_1\right)\left(C_3^2 p_2^2 q_2\right) + p_1^3 p_2^3 = 0{,}321$;
б) $\left(C_3^1 p_1 q_1^2\right) q_2^3 + C_3^2 p_1^2 q_1 \left(q_2^3 + C_3^1 p_2 q_2^2\right) + p_1^3 \left(1 - p_2^3\right) = 0{,}243$.
2.25. а) $P_{6,10}= C_{10}^6 p^6 q^4 = 0{,}251$; б) $P_{120,200} = 0{,}0576$.
2.26. а) при p = 0,9 $P_{200}(180 \leq m \leq 200)$=0,5; б) при p = 0,1 $P_{10}(m \leq 2) = P_{0,10} + P_{1,10} + P_{2,10} = 0{,}930$.
2.27. а) $P_{180,400}$=0,0054; б) $P_{400}(180 \leq m \leq 400) = 0{,}977$.
2.28. а) $P_{1800}(300 \leq m \leq 1800) = 0{,}998$; б) $P_{1800}(300 \leq m \leq 400) = 0{,}906$.
2.29. $n = 55$.
2.30. $m_0 = 8$; $P_{8,800}=P_8(8)=0{,}1396$.
2.31. а) $0{,}88 \leq w \leq 0{,}92$; б) $P_{900}(0{,}08 \leq w \leq 0{,}11) = 0{,}8186$.
2.32. $n = t^2 pq/\Delta^2 = 681$, где $\Phi(t) = 0{,}991$.
2.33. $n = t^2 pq/\Delta^2 = 1089$, где $\Phi(t) = 0{,}996$.
2.34. а) $P_{10000}(100 \leq m \leq 10000) \approx 0$; $P_{10000}(50 \leq m \leq 10000)$=0,750.
2.35. $(1-P_{0,10})(1-P_{0,8}-P_{1,8}) = 0{,}3235$.
2.36. а) $P_6(2;2;2)$=0,081; б) $P_6(0;6;0) + P_6(1;4;1) + P_6(2;2;2) +$ $+ P_6(3;0;3) = 0{,}213$.

Глава 3

3.25. $X=\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0{,}4096 & 0{,}4096 & 0{,}1536 & 0{,}0256 & 0{,}0016 \end{pmatrix}$.
3.26. $X=\begin{pmatrix} 0 & 5 & 10 & 15 \\ 0{,}512 & 0{,}384 & 0{,}096 & 0{,}008 \end{pmatrix}$, $M(X)=5np=3$.

3.27. $X=\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 0{,}59049 & 0{,}32805 & 0{,}07290 & 0{,}00810 & 0{,}00045 & 0{,}00001 \end{pmatrix}$.
$M(X) = np = 0{,}5$; $D(X) = npq = 0{,}45$.

3.28. $X=\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 0{,}00001 & 0{,}00045 & 0{,}00810 & 0{,}07290 & 0{,}32805 & 0{,}59049 \end{pmatrix}$,
$M(X) = np = 4{,}5$; $D(X) = npq = 0{,}45$, где X (тыс. руб.).

3.29. $X=\begin{pmatrix} 0 & 1 & 2 & 3 \\ 27/64 & 27/64 & 9/64 & 1/64 \end{pmatrix}$, $M(X) = np = 3/4$;
$D(X) = npq = 9/16$.

3.30. $X=\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0{,}6561 & 0{,}2916 & 0{,}0486 & 0{,}0036 & 0{,}0001 \end{pmatrix}$, $M(X) = np = 0{,}4$;
$D(X) = npq = 0{,}36$.

3.31. $X=\begin{pmatrix} 0 & 1 & 2 & 3 \\ 0{,}006 & 0{,}092 & 0{,}398 & 0{,}504 \end{pmatrix}$, $M(X) = 2{,}4$; $D(X) = 0{,}46$.

3.32. $X=\begin{pmatrix} 0 & 1 & 2 & 3 \\ 0{,}024 & 0{,}188 & 0{,}452 & 0{,}336 \end{pmatrix}$, $M(X)=2{,}1$; $D(X) = 0{,}61$;
$\sigma(X) = 0{,}78$.

3.33. $X=\begin{pmatrix} 0 & 1 & 2 \\ 0{,}06 & 0{,}38 & 0{,}56 \end{pmatrix}$, $M(X)=1{,}5$; $D(X)=0{,}37$;
$F(x) = \{0$ при $x \le 0$; $0{,}06$ при $0 < x \le 1$; $0{,}44$ при $1 < x \le 2$;
1 при $x > 2\}$.

3.34. $X=\begin{pmatrix} 0 & 1 & 2 & 3 \\ 0{,}06 & 0{,}29 & 0{,}44 & 0{,}21 \end{pmatrix}$, $M(X)=1{,}8$; $D(X)=0{,}7$.

3.35. $F(x) = \{0$ при $x \le 2$; $0{,}3$ при $2 < x \le 4$; 1 при $x > 4\}$.

3.36. $X=\begin{pmatrix} 0 & 1 & 2 \\ 0{,}3 & 0{,}6 & 0{,}1 \end{pmatrix}$; $F(x) = \{0$ при $x \le 0$; $0{,}3$ при $0 < x \le 1$;
$0{,}9$ при $1 < x \le 2$; 1 при $x > 2\}$.

3.37. $X=\begin{pmatrix} 0 & 1 & 2 & 3 \\ 0{,}17 & 0{,}5 & 0{,}3 & 0{,}03 \end{pmatrix}$.

3.38. $X=\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 0{,}04196 & 0{,}25175 & 0{,}41958 & 0{,}23976 & 0{,}04496 & 0{,}00200 \end{pmatrix}$.

3.39. $X=\begin{pmatrix} 0 & 1 & 2 & 3 \\ 1/14 & 3/7 & 3/7 & 1/14 \end{pmatrix}$;
$F(x) = \{0$ при $x \le 0$; $1/14$ при $0 < x \le 1$;
$1/2$ при $1 < x \le 2$; $13/14$ при $2 < x \le 3$; 1 при $x > 3\}$.

3.40. $X=\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0{,}3 & 0{,}21 & 0{,}147 & 0{,}343 \end{pmatrix}$, $M(X)=2{,}533$; $D(X)=1{,}5349$.

3.41. $X=\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1/3 & 2/9 & 4/27 & 8/27 \end{pmatrix}$.

3.42. $X=\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0{,}4 & 0{,}24 & 0{,}144 & 0{,}0864 & 0{,}1296 \end{pmatrix}$; $M(X)=2{,}3056$;
$D(X)=1{,}9626$.

3.43. $X=\begin{pmatrix} 1 & 2 & 3 \\ 0{,}10 & 0{,}18 & 0{,}72 \end{pmatrix}$; $M(X)=2{,}62$.

3.44. а) $X=\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0{,}6 & 0{,}2 & 0{,}08 & 0{,}12 \end{pmatrix}$; б) $M(X)=1{,}72$; $D(X) = 1{,}0816$.

3.45. $X=\begin{pmatrix} 1 & 2 & 3 & 4 \\ 7/10 & 7/30 & 7/120 & 1/120 \end{pmatrix}$; $M(X)=1{,}375$; $D(X)=0{,}401$.

3.46. $X=\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 \end{pmatrix}$; $M(X)=2{,}5$; $D(X)=1{,}25$; $\sigma(X) = 1{,}12$.

3.47. $X=\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0{,}2 & 0{,}2 & 0{,}2 & 0{,}2 & 0{,}2 \end{pmatrix}$; $M(X)=3$;
$F(x) = \{0$ при $x \le 1$; $0{,}2$ при $1 < x \le 2$; $0{,}4$ при $2 < x \le 3$;
$0{,}6$ при $3 < x \le 4$; $0{,}8$ при $4 < x \le 5$; 1 при $x > 5\}$.

3.48. $X=\begin{pmatrix} 1 & 2 & 3 \\ 0{,}3 & 0{,}4 & 0{,}3 \end{pmatrix}$; б) $M(X)=2$; $D(X)=0{,}6$.

3.49. $Z = 3X - 2Y = \begin{pmatrix} -6 & -4 & -3 & -1 & 3 & 5 \\ 0{,}12 & 0{,}08 & 0{,}30 & 0{,}20 & 0{,}18 & 0{,}12 \end{pmatrix}$; $M(X)=1{,}4$;
$M(Y)=2{,}6$; $M(Z)=-1{,}0$; $D(X)=1{,}24$; $D(Y)=0{,}24$; $D(Z)=12{,}12$.

3.50. а) $Z=X+Y=\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0{,}05 & 0{,}30 & 0{,}20 & 0{,}30 & 015 \end{pmatrix}$; б) $M(X)=1{,}2$;
$M(Y)=1{,}0$; $M(Z)=2{,}2$.

3.51. $Z=X+Y=\begin{pmatrix}-1 & 0 & 1 & 2 & 3\\ 0{,}016 & 0{,}176 & 0{,}436 & 0{,}336 & 0{,}036\end{pmatrix}$, $M(Z)=1{,}2$.

3.52. $Z=2X=\begin{pmatrix}2 & 4 & 8\\ 0{,}2 & 0{,}3 & 0{,}5\end{pmatrix}$; $U=X+Y=$

$=\begin{pmatrix}2 & 3 & 4 & 5 & 6 & 8\\ 0{,}04 & 0{,}12 & 0{,}09 & 0{,}20 & 0{,}30 & 0{,}25\end{pmatrix}$; $M(X)=M(Y)=2{,}8$; $M(Z)=$ $=M(2X)=5{,}6$; $M(U)=M(X+Y)=5{,}6$.

3.53. $Z=X^2=\begin{pmatrix}1 & 4 & 16\\ 0{,}2 & 0{,}3 & 0{,}5\end{pmatrix}$; $U=XY=\begin{pmatrix}1 & 2 & 4 & 8 & 16\\ 0{,}04 & 0{,}12 & 0{,}29 & 0{,}30 & 0{,}25\end{pmatrix}$; $M(X)=M(Y)=2{,}8$; $M(U)=M(XY)=7{,}84=2{,}8^2$.

3.54. $Z=\begin{pmatrix}0 & 1 & 2 & 3 & 4\\ 0{,}0144 & 0{,}1104 & 0{,}3124 & 0{,}3864 & 0{,}1764\end{pmatrix}$; б) $M(Z)=np_1+np_2=$ $=2{,}6$; $D(Z)=np_1q_1+np_2q_2=0{,}9$.

3.55. $Z=\begin{pmatrix}1 & 2 & 3 & 4\\ 1/6 & 1/3 & 1/3 & 1/6\end{pmatrix}$.

3.56. $X=\begin{pmatrix}25 & 27 & 50 & 54\\ 0{,}28 & 0{,}42 & 0{,}12 & 0{,}18\end{pmatrix}$, где X (тыс. руб.).

3.57. $X=\begin{pmatrix}-25 & -10 & -4 & +11\\ 0{,}03 & 0{,}12 & 0{,}17 & 0{,}68\end{pmatrix}$; $M(X)=4{,}85$ (тыс. руб.).

3.58. $P_A(B)=P(AB)/P(A)=0{,}8$, где $A=(X>2)$, $B=(X<5)$.

3.59. а) $P(X_1+X_2)=5/8$; б) $P_A(B)=P(AB)/P(A)=1{,}5$, где $A=[(X_1+X_2)>2]$, $B=(X_1=3)$.

3.60. а) $C=1/55$; б) $P[|X-2|\le 1]=P(X=1)+P(X=2)+P(X=3)=14/55$.

3.61. а) $C=0{,}5$; б) $P(X\le 3)=P(X=0)+P(X=1)+P(X=2)+$ $+P(X=3)=0{,}9375$.

3.62. $P(0\le X\le 2)=F(2)-F(0)=0{,}5$.

3.63. а) $P(2\le X<4)=F(4)-F(2)=1/4$; б) $P(2\le X<6)=1$;
в) $P(3\le X\le 6)=F(6)-F(3)=15/16$;
г) $P(6\le X\le 6)=P(X=6)=0$.

3.64. $a=-1/9$; $b=8/9$; $c=-7/9$; $P(2\le X\le 3)=F(3)-F(2)=1/3$.

3.65. $C=1$; $M(X)=2$; $D(X)=2$.

3.66. а) $\varphi(x)=\{0$ при $x\le 0$ и при $x>1$; $2x$ при $0<x\le 1\}$;
б) $M(X)=0{,}6667$; в) $D(X)=0{,}0556$; г) $P(X=0{,}5)=0$; $P(X<0{,}5)=0{,}25$; $P(0{,}5\le X\le 1)=0{,}75$.

3.67. а) $Mo(X)=1$; $Me(X)=0{,}707$; б) $x_{0,4}=0{,}632$; $x_{0,8}=0{,}894$.

3.68. а) $A=\mu_3/\sigma^3=-0{,}566$; б) $E=\mu_4/\sigma^4-3=-0{,}6$.

3.69. а) $A=1/\pi$; б) $F(x)=(1/\pi)\operatorname{arctg}x+1/3$; в) $P(-1\le X\le 1)=0{,}5$. $M(X)$ и $D(X)$ не существуют, так как выражающие их интегралы расходятся.

3.70. а) $A=\lambda/2$; б) $F(x)=\{0{,}5e^{+\lambda x}$ при $x\le 0$; $1-0{,}5\ e^{-\lambda x}$ при $x>0\}$; в) $M(X)=0$, $D(X)=2/\lambda^2$.

3.71. а) $\varphi(x)=\{(2/C)(1-x/c)$ при $0<x\le c$; 0 при $x>c\}$;
б) $M(X)=c/3$; $D(X)=c^2/18$; $\mu_3(X)=c^3/135$; $P(c/2\le X\le c)=$ $=1/4$.

3.72. а) $\varphi(x)=\{(1/3)(1-|x|/3)$ при $-3\le x\le 3$; 0 при $x<-3$ или $x>3\}$; $F(x)=\{0$ при $x\le -3$; $(1/18)\ (x+3)^2$ при $-3<x\le 0$; $(1/18)(6x-x^2+9)$ при $0<x\le 3$; 1 при $x>3\}$; б) $M(X)=0$, $D(X)=1{,}5$; $\mu_3(X)=0$; в) $P(-3/2\le X\le 3)=7/8$.

Глава 4

4.11. $P(X=m)=C_{19}^m 0{,}1^m\cdot 0{,}9^{19-m}$, $m=0,1,2,\ldots,19$; $M(X)=np=$ $1{,}9$; $D(X)=npq=1{,}71$; $Mo(X)_1=1$, $Mo(X)_2=2$.

4.12. $M_{m/n}=0{,}1$; $D(m/n)=0{,}00473$.

4.13. $F(x)=\{0$ при $x<0$; $\sum_{m=0}^{k}C_n^m p^m q^{n-m}$ при $k<x<k+1$; 1 при $x>n\}$.

4.14. а) $P(X=m)=2^m e^{-2}/m!$, $m=0,1,2,\ldots$; б) $M(X)=\lambda=2$, $D(X)=\lambda=2$; в) $P(X\ge 1)=1-P(X=0)=0{,}865$.

4.15. а) $P(X=m)=0{,}05\cdot 0{,}95^{m-1}$, $m=1,2,\ldots$; б) $M(X)=1/p=20$, $D(X)=q/p^2=380$; в) $P(X\ge 5)=1-P(X<5)=1-$ $-[P(X=1)+P(X=2)+P(X=3)+P(X=4)]=0{,}8145$.

4.16. а) $P(X=m)=C_7^m C_{13}^{5-m}/C_{20}^5$, $m=0,1,2,\ldots,5$; б) $M(X)=nM/N=$ $=1{,}75$; $D(X)=nM(1-M/N)(1-n/N)/(N-1)=0{,}898$;
в) $P(X=0)=0{,}083$.

4.17. 1) $M(X)=0{,}1$; $D(X)=0{,}00333$; $\sigma(X)=0{,}0577$;

2.а) $P(0 < X < 0{,}04) + P(0{,}16 < X < 0{,}20) = 0{,}4$;

б) $P(0{,}05 < X < 0{,}15) = 0{,}5$.

4.18. а) $\varphi(x) = \{0$ при $x < 0$; $0{,}0125e^{-0{,}0125x}$ при $x \geq 0\}$; $F(x) = \{0$ при $x < 0$; $1 - e^{-0{,}0125x}$ при $x \geq 0\}$, б) $P = 1 - F(100) = 0{,}286$.

4.19. 1.а) $P(X \leq 15{,}3) = F(15{,}3) = 0{,}9332$; б) $P(X \geq 15{,}4) = 1 - F(15{,}4) = 0{,}0228$; в) $P(14{,}9 \leq X \leq 15{,}3) = 0{,}6246$;
2) $14{,}4 \leq X \leq 15{,}6$.

4.20. а) $M(X) \approx 98$ (ден. ед.), $\sigma(X) \approx 12$ (ден. ед.); б) $P(83 \leq X \leq 96) = 0{,}461$; в) $\Delta = t\sigma = 23{,}52$ (ден. ед.), где $\Phi(t) = 0{,}95$.

4.21. а) $P(X \leq 470) = 0{,}050$; б) $P(500 \leq X \leq 550) = 0{,}609$;
в) $P(X > 550) = 0{,}341$; г) $P(|X - 540| \leq 30) = 0{,}781$.

4.22. $\sigma \approx 10$, $P(35 \leq X \leq 40) = 0{,}09$; б) $P(30 \leq X \leq 35) \approx 0{,}15$.

4.23. а) Из интервала (1,2), так как $P(1<X<2) = 0{,}3414$,
а $P(2 < X < 6) = 0{,}1586$.

4.24. $a \approx 15{,}2$; $\sigma \approx 3{,}1$.

4.25. Параметры $a \approx 28{,}6$; $\sigma \approx 25{,}5$; $P(35 \leq X \leq 45) \approx 0{,}29$.

4.26. а) Параметры $a \approx 781$; $\sigma \approx 0{,}703$; $P(X \geq 1000) = 0{,}363$;

б) $P(X < 500) = 0{,}264$.

4.27. $F_N(x) = 0{,}5 + 0{,}5\phi((x - 150/50)$.

4.28. $\varphi_N(x) = 1/(1{,}48\sqrt{2\pi})e^{-x^2/4{,}38}$; $F_N(x) = 0{,}5 + 0{,}5\phi(x/1{,}48)$.

4.29. $\alpha = a - \sigma\sqrt{3}$; $\beta = a + \sigma\sqrt{3}$ /

4.30. $\sigma = \sqrt{3/2\ln 2} \approx 1{,}47$. **4.31.** а) $\varphi(x) = \{4e^{-4x}$ при $x \geq 0$; 0 при $x < 0\}$; б) $\varphi(x) = \{4e^{2x}(1 - e^{-2x})$ при $x \geq 0$; 0 при $x < 0\}$.

Глава 5

5.10. $X = \begin{pmatrix} -1 & 0 & 1 \\ 0{,}15 & 0{,}50 & 0{,}35 \end{pmatrix}$, $Y = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0{,}11 & 0{,}33 & 0{,}40 & 0{,}16 \end{pmatrix}$;

б) $X_{Y=2} = \begin{pmatrix} -1 & 0 & 1 \\ 0{,}225 & 0{,}4 & 0{,}375 \end{pmatrix}$, $Y_{X=0} = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0{,}08 & 0{,}40 & 0{,}32 & 0{,}20 \end{pmatrix}$;

в) $P(Y > X) = 0{,}81$.

5.11. а) $\varphi(x,y) = \{1/900$ при $0 \leq x \leq 30$, $0 \leq y \leq 30$; 0 при $x < 0$, или $x > 30$, или $y < 0$, или $y > 30\}$; $F(x,y) = \{0$ при $x < 0$ или $y < 0$; $(1/900)\,xy$ при $0 \leq x \leq 30$, $0 \leq y \leq 30$ $(1/30)x$ при $0 \leq x \leq 30$, $y > 30$; $(1/30)y$ при $x > 30$, $0 \leq y \leq 30$; 1 при $x > 30$, $y > 30\}$; б) $\varphi_1(x) = \{1/30$ при $0 \leq x \leq 30$; 0 при $x < 0$ и $x > 30\}$, $\varphi_2(y) = \{1/30$ при $0 \leq y \leq 30$; при $y < 0$ и $y > 30\}$; $F_1(x) = \{0$ при $x < 0$; $(1/30)x$ при $0 \leq x \leq 30$; 1 при $x > 30\}$; $F_2(y) = \{0$ при $y < 0$; $(1/30)y$ при $0 \leq y \leq 30$; 1 при $y > 30\}$; в) X и Y независимы;

г) $P[(0 \leq X < y)(0 \leq Y \leq 30)] = 0{,}5$; $P[(y < X \leq 30)(0 \leq Y \leq 30)] = 0{,}5$.

5.12. а) $\varphi(x,y) = \{1/2$ при $(x,y) \in R$; 0 при $(x,y) \notin R\}$;

б) $\varphi_1(x) = \{0$ при $|x| > 1$; $1 + x$ при $-1 < x < 0$; $1 - x$ при $0 < x < 1\}$;

$\varphi_2(x) = \{0$ при $|y| > 1$; $1 + y$ при $-1 < y < 0$; $1 - y$ при $0 < y < 1\}$; в) $\varphi_y(x) = \{0$ при $|x| > 1 - |y|$; $1/[2(1 - |y|)]$ при $-(1 - |y|) < x < (1 - |y|)\}$, $\varphi_x(y) = \{0$ при $|y| > 1 - |x|$;

$1/[2(1 - |x|)]$ при $(1 - |x|) < y < (1 - |x|)$; г) X и Y зависимы.

5.13. $\varphi(x,y) = \{0$ при $x < 0$ или $y < 0$; $10e^{-(5x+2y)}$ при $x > 0$, $y > 0\}$; $F(x,y) = \{0$ при $x < 0$ или $y < 0$; $(1 - e^{-5x})(1 - e^{-2y})$ при $x > 0$, $y > 0\}$.

5.14. а) $K_{xy} = -0{,}012$; $\rho = -0{,}02$; б) X и Y коррелированы.

5.15. $K_{xy} = 0$; $\rho = 0$; б) X и Y некоррелированы.

5.16. $K_{xy} = 0$; $\rho = 0$; б) X и Y некоррелированы.

5.17. $g(y) = \{0$ при $y \leq 0$; $e^{-\sqrt{y}}/2\sqrt{y}$ при $y > 0\}$, $M(Y) = 2$.

5.18. $N(0;2)$, т.е. $\varphi(z) = (1/(2\sqrt{\pi}))e^{-z^2/4}$.

5.19.

а) (X, Y):

y_i / x_i	0	1
0	1/3	1/6
0	1/3	1/6

; $X = \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}$; $Y = \begin{pmatrix} 0 & 1 \\ 2/3 & 1/3 \end{pmatrix}$.

б) $X_{Y=0} = \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}$, $X_{Y=1} = \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}$; $Y_{X=0} = \begin{pmatrix} 0 & 1 \\ 2/3 & 1/3 \end{pmatrix}$,

$Y_{X=1} = \begin{pmatrix} 0 & 1 \\ 2/3 & 1/3 \end{pmatrix}$.

5.20. $\varphi(x,y)=\{1$ при $0\le x\le 1$, $0\le y\le 1$; 0 — в остальных случаях$\}$; $F(x,y)=\{0$ при $x\le 0$ или $y\le 0$; xy при $0<x\le 1$ и $0<y<1$; x при $0<x\le 1$ и $y>1$; y при $x>1$ и $0<y\le 1$; 1 при $x>1$ и $y>1\}$.

5.21. $\varphi(x,y)$ $\{(3/\pi)(1-\sqrt{x^2+y^2})$ при $x^2+y^2\le 1$; 0 при $x^2+y^2>1\}$; $\varphi_1(x,y)$ $\{(3/\pi)$ $[\sqrt{1-x^2}-x^2\ln((1+\sqrt{1-x^2})/|x|)]$ при $|x|<1$; 0 при $|x|>1\}$; $\varphi_2(x,y)$ $\{(3/\pi)$ $[\sqrt{1-y^2}-y^2\ln((1+\sqrt{1-y^2})/|y|)]$ при $|y|<1$; 0 при $|y|>1\}$; $\varphi_x(y)=\{\left(1-\sqrt{x^2-y^2}\right)/\left[\sqrt{1-x^2}-x^2\ln\left(1+\sqrt{1-x^2}\right)/|x|\right]$; при $|y|<\sqrt{1-x^2}$; $|x|<1$; 0 при $|y|<\sqrt{1-x^2}$, $|x|<1\}$; $\varphi_y(x)=$ $=\{(1-\sqrt{x^2+y^2})/[\sqrt{1-y^2}-y^2\ln((1+\sqrt{1-y^2})/|y|)$ при $|x|<\sqrt{1-y^2})$ $|y|<1$; 0 при $|x|<\sqrt{1-y^2}$ $|y|<1\}$; X и Y зависимы, так как $\varphi_y(x)\ne\varphi_1(x)$, $\varphi_x(y)\ne\varphi_2(y)$, но не коррелированы, так как $\rho=0$.

5.22. а) $A=1/\pi^2$; б) $P[(-1\le X\le 1)(-1\le Y\le 1)]=1/4$; $\varphi_1(x)=1/[\pi(1+x^2)]$, $\varphi_2(x)=1/[\pi(1+y^2)]$; X и Y независимы, так как $\varphi(x,y)=\varphi_1(x)\varphi_2(y)$. **5.23.** а) $C=12/11$; б) $\varphi_1(x)=12(x^2+x/2+1/2)/11$, $\varphi_1(y)=12(y^2+y/2+1/2)/11$; в) $\varphi_y(x)=(x^2+x/y+y^2)/(y^2+y/2+1/2)$, $\varphi_x(y)=(x^2+x/y+y^2)/(x^2+x/2+1/2)$; г) $a_x=a_y=7/11$; $D(X)=D(Y)=257/3630$; $\rho=-40/257\approx -0{,}156$. **5.24.** $\varphi(x,y)=\{2^{-x-y}\cdot\ln^2 2$ при $x\ge 0$, $y\ge 0$; 0 при $x<0$ или $y<0\}$; $P[(1\le X<2)(3\le Y\le 5)]=3/128$. 5.25. $F(x,y)=[(1/\pi)\mathrm{arctg}(x/4)+0{,}5]\,[(1/\pi)\mathrm{arctg}(y/5)+0{,}5]$. **5.26.** $\varphi(x,y)=\{(1/\pi)e^{-x^2}$ при $0\le y\le 1$; 0 при $y<0$ или $y>1\}$; $F(x,y)=\{0$ при $y\le 0$; $y(0{,}5+0{,}5\,\varphi(x\sqrt{2})$ при $0<y\le 1$; $0{,}5+0{,}5\,\varphi(x\sqrt{2})$ при $y>1\}$. **5.27.** $a_x=2$, $a_y=-3$, $\sigma_x^2=\sigma_y^2=1$; $\rho=0{,}6$. **5.28.** а) 0,25; б) 0,5; в) $0{,}25[\phi(3)-\phi(1)]\,[\phi(2{,}5)-\phi(0{,}5)]\approx 0{,}0476$. **5.29.** а) $g(y)=(1/y)\,[\ln(1/y)]$ $(0<y<1)$; б) $g(y)=e^y\,\varphi(e^y)$; в) $(1/3\sqrt[3]{y^2}\,\varphi\sqrt[3]{y})$, $(0<y<\infty)$; г) $g(y)=(1/(2y\sqrt{y})\varphi(1/\sqrt{y})$ $(0<y<\infty)$; д) $g(y)=2y\varphi(y^2)$ $(0<y<\infty)$. **5.30.** $g(y)=\{1/(\pi 1/\left(\pi\sqrt{1-y^2}\right)$ при $-1<y<1$; 0 при $y<-1$ или $y>1\}$. 5.31. $g(y)=0{,}5/\sqrt{y}$ $(0<y\le 1)$. **5.32.** $g(y)=1/[\pi(1+y^2)]$, $(-\infty<y<+\infty)$. **5.33.** $a_y=2{,}4$, $D(Y)=1{,}99$.

5.34. а) $a_y=5$; $D(Y)=36$; б) $K_{xy}=5$, $\rho=-1$. **5.35.** $a_y=(\pi^2-8)/4$. **5.36.** $a_y=\ln 2$; $D(Y)=0{,}5-\ln^2 2$. **5.37.** $a_y=0{,}5$; $D(Y)=1/12$. **5.38.** $a_y=2/3$; $D(Y)=1/18$. **5.39.** $a_y=-14$. **5.40.** $M(X-Y)=0$, $D(X-Y)=13/3$; б) $M(X^2)=5$, $M(Y^2)=64/45$. **5.41.** $F(z)=\{0$ при $z\le 0$; $z^2/8$ при $0<z\le 2$; $1-(4-z)2/8$ при $2<z\le 4$; 1 при $z>4\}$; $\varphi(z)=\{0$ при $z\le 0$, $z/4$ при $0<z\le 2$; $1-z/4$ при $2<z\le 4$; 0 при $z>4\}$. **5.42.** $\varphi_N(z)=(1/2\pi)e^{-z^2/4}$; $a_z=0$, $\sigma_z^2=2$.

Глава 6

6.9. $P\le 0{,}1$.
6.10. а) $P\ge 0{,}5$; б) $P\le 2/3$.
6.11. $P\ge 0{,}9856$; $P\approx 1$.
6.12. $P\ge 0{,}264$.
6.13. $P\ge 0{,}6$; $P\approx 0{,}8859$.
6.14. $P\ge 0{,}75$.
6.15. $P\ge 0{,}996$.
6.16. $P\ge 0{,}9344$.
6.17. $0{,}2\le w\le 0{,}4$.
6.18. $P\ge 0{,}91$.
6.19. а) $n\le 595$; б) $n\le 121$.
6.20. $n\ge 16000$; $n\approx 4330$.
6.21. $P\ge 0{,}929$.
6.22. $n\ge 3333$.

Глава 7

7.10. $a_x(t)=ae^{-t}$; $D_x(t)=\sigma^2e^{-2t}$; $K_x(t_1,t_2)=\sigma^2e^{-(t_1+t_2)}$, $r_x(t_1,t_2)\equiv 1$.
7.13. а) $P_y(2)=0{,}134$; б) $P(X\ge 1)=1-P_0(2)=0{,}9975$; в) $P_0(2)=0{,}0025$.
7.14. $p_0=8/87$, $p_1=46/87$, $p_2=10/87$, $p_3=23/87$ (для графа на рис. 7.11); $p_0=10/17$, $p_1=4/17$, $p_2=3/17$ (для графа на рис. 7.12).
7.15. $p_0=Q=4/7$; $p_1=P_{\text{отк}}=3/7$; $A=6/7$ (машин в час).
7.16. $p_0=0{,}473$, $p_1=0{,}355$, $p_2=0{,}133$, $p_3=0{,}033$, $p_4=P_{\text{отк}}=0{,}006$; $Q=0{,}994$, $A=1{,}491$ (машин в час); $n=3$. **7.17.** $p_0=Q=0{,}455$, $p_1=P_{\text{отк}}=0{,}545$, $A=0{,}182$ разг./мин; номинальная пропускная способность $A_{\text{ном}}=0{,}333$ разг./мин. почти вдвое больше. **7.18.** При $n=2$ $p_0=0{,}107$, $p_2=0{,}550$, $Q=1-p_2=0{,}450$, $A=1{,}8$ заявки/ч, доход $D=7{,}20$ ден. ед./ч; при $n=3$ $p_0'=0{,}0677$, $p_3'=0{,}371$, $Q'=1-p_3'=0{,}629$, $A'=2{,}52$ заявки/ч, доход $D'=10{,}08$ ден. ед./ч; увеличение числа каналов с $n=2$ до $n=3$ выгодно, так как увеличение дохода ($D'-D=2{,}88$ ден. ед./ч) превосходит увеличение затрат (2 ден. ед./ч).

Глава 8

8.10. 2.а) $\bar{x}=1{,}535$; б) $\tilde{Me}=1$; $\tilde{Mo}=0$; в) $s^2=3{,}378$; $s=1{,}838$; $\tilde{\nu}=119{,}7(\%)$; г) $\tilde{\nu}_1=\bar{x}=1{,}535$; $\tilde{\nu}_2=5{,}735$; $\tilde{\nu}_3=30{,}38$; $\tilde{\nu}_4=201{,}605$; $\tilde{\mu}_1=0$; $\tilde{\mu}_2=3{,}378$; $\tilde{\mu}_3=11{,}2039$; $\tilde{\mu}_4=79{,}494113$; д) $\tilde{A}=1{,}80$; $\tilde{E}=3{,}97$.

8.11. 2.а) $\bar{x}=1653$ (руб.); б) $\tilde{Me}=1663$ (руб.); $\tilde{Mo}=1665$ (руб.), в) $s^2=44559$; $s=667{,}5$ (руб.); $v=40{,}4(\%)$; г) $\tilde{\nu}_1=\bar{x}=1653$; $\tilde{\nu}_2=317800$.

8.12. 2,а) $\bar{x}=15{,}6$ (ц); б) $\tilde{Me}=15{,}4$ (ц); $\tilde{Mo}=14{,}9$ (ц); в) $s^2=19{,}0$; $s=4{,}36$(ц); $\tilde{\nu}=27{,}9(\%)$; г) $\tilde{\nu}_1=\bar{x}=15{,}6$; $\tilde{\nu}_2=262{,}36$; $\tilde{\nu}_3=4685{,}76$; $\tilde{\nu}_4=87874{,}12$; $\mu_1=0$; $\mu_2=s^2=19{,}0$; $\mu_3=0{,}144$; $\mu_4=898{,}0048$.

8.13. $s^2=418{,}42$; $s_1^2=592{,}05$; $s_2^2=113{,}43$; $\overline{s_i^2}=400{,}60$; $\delta^2=17{,}82$; $s^2=\overline{s_i^2}+\delta^2$.

Глава 9

9.19. 1.а) $P=\Phi(2{,}19)=0{,}9715$; б) $1599{,}9\le\bar{x}_0\le1706{,}1$ (руб.); 2) $n'=1329$.

9.20. 1.а) $P=\Phi(2{,}13)=0{,}967$; б) $1598{,}5\le\bar{x}_0\le1707{,}5$ (руб.); г) $n'=1352$.

9.21. 1.а) $P=\Phi(1{,}39)=0{,}835$; б) $0{,}00412\le w\le0{,}00748$; г) $n'=1605$; 3) при $[w(1-w)]_{max}=0{,}25$ $P\ge\Phi(0{,}65)=0{,}484$; $n'\le1605$.

9.22. 1.а) $P=\Phi(1{,}35)=0{,}823$; б) $0{,}0408\le w\le0{,}0752$; 2) $n'=1658$; 3) при $[w(1-w)]_{max}=0{,}25$ $P\ge\Phi(0{,}63)=0{,}471$; $n'\le165$.

9.23. $P=\Phi(0{,}71)=0{,}522$.

9.24. 1.а) $\Delta=0{,}058$ (ед/ч); б) $\Delta'=0{,}060$(ед./ч); 2) $n=347$ при повторной выборке и $n'=268$ при бесповторной выборке.

9.25. $0{,}8792\le w\le0{,}9208$ при повторной выборке и $0{,}8803\le w\le0{,}9197$ при бесповторной выборке.

9.26. От 27,68 до 32,32%; $n=13978$; $n\le16641$ (при отсутствии данных).

9.27. а) $n=756$; $n'=691$; б) $n=484$; $n'=456$.

9.28. $0{,}16\le p\le0{,}37$, учитывая точную формулу для σ_w и нормальный закон распределения w; $0{,}14\le p\le0{,}36$, учитывая приближенную формулу для σ_w и нормальный закон w.

9.29. $0{,}04\le p\le0{,}44$ (используя рис. 9.3).

9.30. а) $P=0{,}718$; б) $15{,}71\le x_0\le16{,}37(\%)$.

9.31. $22{,}9\le\bar{x}_0\le37{,}1$, учитывая, что $t_{0{,}99;8}=3{,}35$.

9.32. $0{,}44\le\sigma\le1{,}06$, учитывая, что $\chi^2_{0{,}025;11}=21{,}9$; $\chi^2_{0{,}975;11}=3{,}82$.

9.33. $0{,}53\le\sigma\le0{,}70$, учитывая, что при аппроксимации χ^2-распределения нормальным получаем $\chi^2_{0{,}925;99}=127{,}9$; $\chi^2_{0{,}975;99}=72{,}9$.

9.34. а) $P=0{,}347$; б) $4{,}40\le\bar{x}_0\le5{,}60$ (ч).

Глава 10

10.15. а) Влияние существенно, так как $t=2{,}82>t_{0{,}95}=1{,}96$ (двусторонний критерий); б) влияние существенно, так как $t=2{,}82>t_{0{,}9}=1{,}64$ (односторонний критерий).

10.16. Новая технология дает значимое уменьшение среднего расхода сырья, так как $t=3{,}38>t_{0{,}9;20}=1{,}72$ (односторонний критерий).

10.17. Утверждение в рекламе противоречит имеющимся данным, так как $t=0{,}18<t_{0{,}9;22}=1{,}72$ (односторонний критерий).

10.18. Значение $x_9=48$ является аномальным, так как $t=2{,}95>t_{0{,}9;7}=1{,}89$ (односторонний критерий).

10.19. Существенных различий нет, так как:
а) $t=1{,}88<t_{0{,}95}=1{,}96$ (двусторонний критерий);
б) $t=1{,}88<t_{0{,}9}=1{,}64$ (односторонний критерий).

10.20. Различия существенны, так как $\chi^2=8{,}12>\chi^2_{0{,}05;3}=7{,}82$ (односторонний критерий).

10.21. Существенных различий нет, так как:
а) $F=1{,}15<F_{0{,}05;8;12}=2{,}85$ (односторонний критерий);
б) $F=1{,}15<F_{0{,}01;8;12}=4{,}50$ и $F=1{,}15>F_{0{,}99;8;12}=0{,}18$, где $F_{0{,}99;8;12}=\dfrac{1}{F_{0{,}01;12;8}}=\dfrac{1}{5{,}67}=0{,}18$ (двусторонний критерий).

10.22. Существенных различий нет, так как $\chi^2 = 1{,}54 < \chi^2_{0{,}05;3} = 7{,}82$ (односторонний критерий). Первый способ существенно «лучшим» признать нельзя.

10.23. а) Отклонение неслучайно, так как $t = 2{,}73 > t_{0{,}9} = 1{,}64$ (односторонний критерий); б) $P = 0{,}5 + 0{,}5\,\Phi(1{,}09) = 0{,}862$.

10.24. а) Отклонение можно считать случайным, так как: а) $t = 1{,}19 < t_{0{,}9;19} = 1{,}73$ (односторонний критерий); б) $P = 0{,}5 - 0{,}5\,\theta\,(0{,}55;19) = 0{,}29$.

10.25. Инвестиционные вложения допустимы, так как:

а) $\chi^2 = 58{,}5 < \chi^2_{0{,}05;51} = 72{,}1$ (односторонний критерий);

б) $\chi^2 = 58{,}5 < \chi^2_{0{,}01;51} = 79{,}8$ (односторонний критерий);

табличные значения критерия χ^2 при $k = 51$ вычисляем с помощью формулы (**9.51**).

10.26. Можно считать, так как $t = 2{,}33 > t_{0{,}9} = 1{,}64$ (односторонний критерий).

10.27. а) Продукция не должна быть забракована, так как $t = 1{,}67 < t_{0{,}95} = 1{,}96$ (двусторонний критерий); б) продукция должна быть забракована, так как $t = 1{,}67 > t_{0{,}9} = 1{,}64$ (односторонний критерий).

10.28. Гипотеза о нормальном распределении с параметрами $a = 1653$, $\sigma^2 = 445591$ не отвергается, так как:

а) $\chi^2 = 2{,}37 < \chi^2_{0{,}05;2} = 5{,}99$; б) $\lambda = 0{,}35 < \lambda_{0{,}05} = 1{,}36$.

10.29. Гипотеза о нормальном распределении с параметрами $a = 15{,}6$, $\sigma^2 = 19{,}0$ не отвергается, так как:

а) $\chi^2 = 1{,}48 < \chi^2_{0{,}05;10} = 18{,}3$; б) $\lambda = 0{,}3 < \lambda_{0{,}05} = 1{,}36$.

10.30. а) Гипотеза о нормальном распределении с параметрами $a = 16{,}04$; $\sigma^2 = 4{,}2384$ отвергается, так как $\chi^2 = 76{,}4 > \chi^2_{0{,}05;3} = 7{,}82$; б) не отвергается, так как $\lambda = 0{,}95 < \lambda_{0{,}05} = 1{,}36$.

10.31. Гипотеза о показательном распределении с параметром $\lambda = 0{,}2$ не отвергается, так как:

а) $\chi^2 = 1{,}29 < \chi^2_{0{,}05;2} = 5{,}99$; б) $\lambda = 0{,}47 < \lambda_{0{,}05} = 1{,}36$.

10.32. Можно считать, так как: а) $\lambda' = 0{,}35 < \lambda_{0{,}05} = 1{,}36$;

б) $\chi^2 = 12{,}0 < \chi^2_{0{,}05;7} = 14{,}1$ (если данные x_i, y_i сгруппировать в одни и те же 8 интервалов).

10.33. Гипотеза о биноминальном законе распределения с параметром $p = 0{,}88$ не отвергается, так как:

а) $\chi^2 = 2{,}71 < \chi^2_{0{,}05;1} = 3{,}84$; б) $\lambda = 0{,}27 < \lambda_{0{,}05} = 1{,}36$.

10.34. Гипотеза о законе распределения Пуассона с параметром $\lambda = 0{,}9$ не отвергается, так как: а) $\chi^2 = 9{,}27 < \chi^2_{0{,}05;4} = 9{,}49$; б) $\lambda = 0{,}88 < \lambda_{0{,}05} = 1{,}36$.

10.35. а) Партия не удовлетворяет гарантии, так как $t = 2{,}95 > t_{0{,}9} = 1{,}64$; б) $P = \frac{1}{2} + \frac{1}{2}\phi\ (1{,}30) = \frac{1}{2} + \frac{1}{2} \cdot 0{,}8064 = 0{,}903$; в) $n = (t_{0{,}9;n-1} + t_{0{,}96;n-1})^2 \cdot 110^2/50^2 = 70$ (найдено подбором n).

Глава 11

11.3. Влияние типа технологии (фактора A) на урожайность значимо, так как $F = s_1^2/s_2^2 = 9{,}35 < F_{0{,}05;4;25} = 2{,}76$.

11.4. Влияние линии (фактора A) на качество облицовочных плиток незначимо, так как $F = s_1^2/s_2^2 = 1{,}29 < F_{0{,}05;3;36} = 2{,}87$.

11.5. Влияние на урожайность: сорта пшеницы (фактора A) значимо, (так как $F = s_1^2/s_3^2 = 5{,}90 > F_{0{,}05;3;12} = 3{,}49$); участков земли — блоков (фактора B) незначимо (так как $F = s_2^2/s_3^2 = 3{,}17 < F_{0{,}05;4;12} = 3{,}26$).

11.6. В рамках модели 1 (с фиксированными уровнями факторов) влияние на производительность труда технологий (фактора A) значимо (так как $F = s_1^2/s_4^2 = 27{,}6 > F_{0{,}05;2;24} = 3{,}40$); хозяйства (фактора B) — незначимо (так как $F = s_2^2/s_4^2 = 2{,}03 < F_{0{,}05;3;24} = 3{,}01$); взаимодействия факторов A и B — значимо (так как $F = s_3^2/s_4^2 = 4{,}41 > F_{0{,}05;6;24} = 2{,}51$).

Глава 12

12.14. $\bar{x}_1 = 3{,}15$; $\bar{x}_2 = 3.75$; $\bar{x}_3 = 6{,}34$, $\bar{x}_4 = 11{,}79$; $\bar{x}_5 = 15{,}75$. $\bar{x}_6 = 19{,}12$ (тыс. руб.); $\bar{y}_1 = 1{,}82$; $\bar{y}_2 = 3{,}15$; $\bar{y}_3 = 5{,}02$; $\bar{y}_4 = 5{,}78$; $\bar{y}_5 = 7{,}0$(кВт · ч); б) $r = 0{,}872$; связь тесная и прямая, r значим, так как $t = 13{,}57 > t_{0{,}95;58} = 2{,}00$; $0{,}784 \le r < 0{,}926$ (с помощью z-преобразования Фишера); в) $\eta_{yx} = 0{,}886$ (значим, так как $F = 41{,}2 > F_{0{,}05;4;55} = 2{,}55$); $\eta_{xy} = 0{,}893$ (значим, так как $F = 42{,}5 > F_{0{,}05;5;54} = 2{,}40$); г) гипотеза о линейной корреляционной зависимости не отвергается, ибо η^2_{yx} близко к r^2 так, что $F = 2{,}10 < F_{0{,}05;3;55} =$ $= 2{,}78$ (или η^2_{xy} близко к r^2 так, что $F = 2{,}47 < F_{0{,}05;4;54} = 2{,}55$); д) $y_x = 0{,}2966x + 1{,}340$; $x_y = 2{,}5609y - 0{,}944$; $0{,}2528 \le \beta_{yx} \le 0{,}3404$; $2{,}1832 \le \beta_{xy} \le 2{,}9386$.

12.15. а) $r = 0{,}969$; связь очень тесная и прямая; r значим (так как $t = 13{,}59 > t_{0{,}95;12} = 2{,}18$); $0{,}901 \le r \le 0{,}991$ (с помощью z-преобразования Фишера); б) $y_x = 0{,}5435x + 7{,}04$; $x_y = 1{,}7276y - 8{,}94$.

12.16. а) $r = 0{,}917$ (значим, так как $t = 9{,}75 > t_{0{,}95;18} = 2{,}10$); б) $\bar{x} = 21{,}25$ (млн руб.). $\bar{y} = 27{,}5$ (млн руб.).

12.17. а) $y_x = 10x + 3838$, $x_y = 0{,}08y - 303{,}8$; б) $y_{x=16{,}5} = 4003$ (усл.ед.). Согласуется, так как гипотеза H_0: $\rho = 0{,}95$ (с помощью z-преобразования Фишера) не отвергается, ибо $t = 1{,}08 < t_{0{,}95} = 1{,}96$.

12.18. $x_y = -1600y + 7832$.

12.19. а) $r_{y1.2} = 0{,}908$ (значим, так как $t = 13{,}0 > t_{0{,}95;36} = 2{,}02$; $r_{y2.1} = -0{,}907$ (значим, так как $|t| = 12{,}9 > t_{0{,}95;36} = 2{,}02$); б) $R_{y.12} = 0{,}908$ (значим, так как $F = 79{,}8 > F_{0{,}05;2;34} = 3{,}28$; в) $R^2_{y.12} = 0{,}824$.

12.20. $\rho = 0{,}678$ (не значим, так как $t = 2{,}06 < t_{0{,}95;5} = 2{,}57$; $\tau = 0{,}524$ (не значим, так как $t = 1{,}65 < t_{0{,}95} = 1{,}96$.

12.21. $W = 0{,}922$ (значим, так как $\chi^2 = 74{,}7 > \chi^2_{0{,}05;9} = 16{,}9$).

Глава 13

13.8. а) $y_x = 0{,}2966x + 1{,}340$; б) $y_{x=10} = 4{,}306$ (кВт · ч); $4{,}075 \le \le M_{x=10}(Y) \le 4{,}537$ (кВт · ч), учитывая, что $t_{0{,}95;58} = 2{,}00$ и $s_{y_{x=10}} = 0{,}115$(кВт · ч); в) $R^2 = 0{,}760$; г) уравнение регрессии значимо, так как $F = 183{,}7 > F_{0{,}05;1;58} = 4{,}00$.

13.9. а) $y = 0{,}5435x + 7{,}04$; б) $R^2 = 0{,}939$; в) уравнение регрессии значимо, так как $F = 184{,}7 > F_{0{,}05;1;12} = 4{,}75$; г) $38{,}23 \le M_{x=60}(Y) \le 41{,}06$ (т/ч), учитывая, что $t_{0{,}95;12} = 2{,}18$; $s_{y_{x=60}} = 0{,}647$ и $34{,}88 \le y^*_{x_0=60} \le 44{,}4$ (т/ч), учитывая, что $t_{0{,}95;12} = 2{,}18$ и $\delta_{y_{x_0=60}} = 2{,}183$ (т/ч).

13.10. $1346 \le M_{x=1300}(Y) \le 1519$ (ден. ед), учитывая, что $t_{0{,}95;28} = 2{,}05$, а $s_{y_{x=1300}} = 42{,}4$ (ден. ед.); $995 \le y^*_{x_0=1300} \le 1870$ (ден. ед.), учитывая, что $t_{0{,}95;28} = 2{,}05$ и $s_{y_{x_0=1300}} = 213{,}4$ (ден. ед.).

13.11. а) $y_x = -50{,}02 + 10{,}309x - 0{,}4237x^2$; б) $R_{yx} = 0{,}990$; $R^2_{yx} = 0{,}980$; в) $y_{x_{max}} = 12{,}7\%$ при $x = 12{,}2$ (т).

13.12. б) $y_x = 8{,}465 - 0{,}5250x$, $s^2 = 1{,}633$; $y_x = -0{,}048 + 32{,}6200/x$, $s = 1{,}660$; в) $\eta_{yx} = 0{,}642$ (значим, так как $F = 7{,}92 > F_{0{,}05;4;45} = 2{,}59$; $r = R_{yx} = -0{,}619$ (значим, так как $|t| = 5{,}46 > t_{0{,}95;48} = 2{,}01$. (регрессия линейная); $R_{yx} = 0{,}631$ (значим, так как $F = 38{,}4 > F_{0{,}05;1;58} = 4{,}04$) (регрессия гиперболическая).

13.13. а) $r_{y_1} = -0{,}851$ (значим, так как $|t| = 7{,}77 > t_{0{,}95;23} = 2{,}07$); $r_{y_2} = 0{,}808$ (значим, так как $t = 6{,}58 > t_{0{,}95;23} = 2{,}07$); $r_{12} = -0{,}519$ (значим, так как $|t| = 2{,}91 > t_{0{,}95;23} = 2{,}07$); $r_{y1,2} = -0{,}857$ (значим, так как $|t| = 6{,}89 > t_{0{,}95;24} = 2{,}06$); $r_{y2,1} = -0{,}815$ (значим, так как $|t| = 0{,}815 > t_{0{,}95;24} = 2{,}06$); $R_{y1.2} = 0{,}953$ (значим, так как $F = 109 > F_{0{,}05;2;22} = 3{,}44$). б) $y_x = 210{,}66 - 1{,}0238x_1 + 8{,}756x_2$; уравнение регрессии значимо по F-критерию; коэффициенты b_1 и b_2 значимы по t-критерию; в) стандартизованные коэффициенты регрессии $b'_1 = 0{,}5904$, $b'_2 = 0{,}5011$; коэффициенты эластичности $E_1 = -0{,}185$; $E_2 = 0{,}150$.

13.14. а) $y_x = 5{,}62 - 0{,}239x_1 + 0{,}774x_2$; б) $R = 0{,}981$; в) $R^2 = 0{,}962$; в) уравнение регрессии значимо, так как $F = 36{,}3 > F_{0{,}05;2;3} = 9{,}55$; г) при $x_1 = 5{,}5$, $x_2 = 6{,}0$ $y_x = 8{,}95(\%)$.

Глава 14

14.8. $\bar{y}_t = 16{,}66$ (ц/га); $s_t = 3{,}62$ (ц/га); $r_1 = 0{,}035$; $r_2 = -0{,}339$.

14.9. $\tilde{y}_t = 14{,}89 + 0{,}322t$. Уравнение тренда незначимо, так как $F = 0{,}56 < F_{0{,}05;1;8} = 5{,}32$.

14.10. При m=3:

t	1	2	3	4	5	6	7	8	9	10
$\hat{y}_t$	–	17,87	15,00	13,37	13,10	17,17	16,87	17,67	17,93	–

При m=5:

t	1	2	3	4	5	6	7	8	9	10
y_t	–	–	15,32	15,32	15,26	14,72	16,92	18,00	–	–

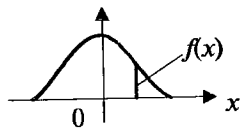
14.11. $\tilde{y}_t = 1094{,}1 + 62{,}19t$; уравнение тренда значимо, так как $F = 32{,}3 > F_{0{,}05;1;6} = 5{,}99$; б) автокорреляция возмущений ряда незначима; в) точечный прогноз $\tilde{y}_9 = 1654$ (ден. ед.); с надежностью 0,95 интервальный прогноз: среднего значения доходов — от 1516 до 1792 (ден. ед.); индивидуального значения доходов — от 1432 до 1876 (ден. ед.).

Приложения

Математико-статистические таблицы

Таблица 1

Значения функции $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$



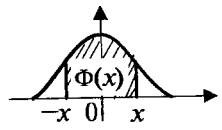
Целые и десятые доли x	Сотые доли x									
	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0,3982	0,3980	0,3977	0,3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
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	3187	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	0,2396	0,2371	0,2347	0,2323	0,2299	0,2275	0,2251	0,2227	0,2203
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	0,0529	0,0519	0,0508	0,0498	0,0488	0,0478	0,0468	0,0459	0,0449
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

Окончание табл. I

x	0	1	2	3	4	5	6	7	8	9
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	0,0043	0,0042	0,0041	0,0039	0,0038	0,0037	0,0036	0,0035	0,0034
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
4,0	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
4,1	0,0001338									
4,5	0,0000160									
5,0	0,0000015									

Таблица II

Значения функции Лапласа $\Phi(x)=\frac{2}{\sqrt{2\pi}}\int_0^x e^{-t^2/2}dt$



Целые и десятые доли x	Сотые доли x									
	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	0,0797	0,0876	0,0955	0,1034	0,1113	0,1192	0,1271	0,1350	0,1428	0,1507
0,2	0,1585	0,1663	0,1741	0,1819	0,1897	0,1974	0,2051	0,2128	0,2205	0,2282
0,3	0,2358	0,2434	0,2510	0,2586	0,2661	0,2737	0,2812	0,2886	0,2960	0,3035
0,4	0,3108	0,3182	0,3255	0,3328	0,3401	0,3473	0,3545	0,3616	0,3688	0,3759
0,5	0,3829	0,3899	0,3969	0,4039	0,4108	0,4177	0,4245	0,4313	0,4381	0,4448
0,6	0,4515	0,4581	0,4647	0,4713	0,4778	0,4843	0,4907	0,4971	0,5035	0,5098
0,7	0,5161	0,5223	0,5285	0,5346	0,5407	0,5467	0,5527	0,5587	0,5646	0,5705
0,8	0,5763	0,5821	0,5878	0,5935	0,5991	0,6047	0,6102	0,6157	0,6211	0,6265
0,9	0,6319	0,6372	0,6424	0,6476	0,6528	0,6579	0,6629	0,6679	0,6729	0,6778
1,0	0,6827	0,6875	0,6923	0,6970	0,7017	0,7063	0,7109	0,7154	0,7199	0,7243
1,1	0,7287	0,7330	0,7373	0,7415	0,7457	0,7499	0,7540	0,7580	0,7620	0,7660
1,2	0,7699	0,7737	0,7775	0,7813	0,7850	0,7887	0,7923	0,7959	0,7984	0,8029
1,3	0,8064	0,8098	0,8132	0,8165	0,8198	0,8230	0,8262	0,8293	0,8324	0,8355
1,4	0,8385	0,8415	0,8444	0,8473	0,8501	0,8529	0,8557	0,8584	0,8611	0,8638
1,5	0,8664	0,8690	0,8715	0,8740	0,8764	0,8789	0,8812	0,8836	0,8859	0,8882
1,6	0,8904	0,8926	0,8948	0,8969	0,8990	0,9011	0,9031	0,9051	0,9070	0,9090
1,7	0,9109	0,9127	0,9146	0,9164	0,9181	0,9199	0,9216	0,9233	0,9249	0,9265
1,8	0,9281	0,9297	0,9312	0,9327	0,9342	0,9357	0,9371	0,9385	0,9392	0,9412
1,9	0,9426	0,9439	0,9451	0,9464	0,9476	0,9488	0,9500	0,9512	0,9523	0,9533
2,0	0,9545	0,9556	0,9566	0,9576	0,9586	0,9596	0,9606	0,9616	0,9625	0,9634
2,1	0,9643	0,9651	0,9660	0,9668	0,9676	0,9684	0,9692	0,9700	0,9707	0,9715
2,2	0,9722	0,9729	0,9736	0,9743	0,9749	0,9756	0,9762	0,9768	0,9774	0,9780
2,3	0,9786	0,9791	0,9797	0,9802	0,9807	0,9812	0,9817	0,9822	0,9827	0,9832
2,4	0,9836	0,9841	0,9845	0,9849	0,9853	0,9857	0,9861	0,9865	0,9869	0,9872
2,5	0,9876	0,9879	0,9883	0,9886	0,9889	0,9892	0,9895	0,9898	0,9901	0,9904
2,6	0,9907	0,9910	0,9912	0,9915	0,9917	0,9920	0,9922	0,9924	0,9926	0,9928
2,7	0,9331	0,9933	0,9935	0,9937	0,9939	0,9940	0,9942	0,9944	0,9946	0,9947
2,8	0,9949	0,9951	0,9952	0,9953	0,9955	0,9956	0,9958	0,9959	0,9960	0,9961
2,9	0,9963	0,9964	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972
3,0	0,9973	0,9974	0,9975	0,9976	0,9976	0,9977	0,9978	0,9979	0,9979	0,9980
3,1	0,9981	0,9981	0,9982	0,9983	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986
3,2	0,9986	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,3	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,4	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995	0,9995
3,5	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997	0,9997
3,6	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998	0,9998	0,9998
3,7	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
4,0	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999

Таблица III

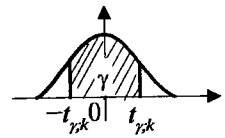
Значения функции Пуассона $P_m(\lambda) = \frac{\lambda^m}{m!} e^{-\lambda}$

m \ λ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679
1	0,0905	0,1637	0,2223	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659	0,3679
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1216	0,1438	0,1647	0,1839
3	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494	0,0613
4	0,0000	0,0001	0,0003	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111	0,0153
5	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0007	0,0012	0,0020	0,0031
6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0005
7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

m \ λ	2,0	3,0	4,0	5,0	6,0	7,0	8,0	9,0	10,0
0	0,1353	0,0498	0,0183	0,0067	0,0025	0,0009	0,0003	0,0001	0,0001
1	0,2707	0,1494	0,0733	0,0337	0,0149	0,0064	0,0027	0,0011	0,0005
2	0,2707	0,2240	0,1465	0,0842	0,0446	0,0223	0,0107	0,0050	0,0023
3	0,1805	0,2240	0,1954	0,1404	0,0892	0,0521	0,0286	0,0150	0,0076
4	0,0902	0,1681	0,1954	0,1755	0,1339	0,0912	0,0572	0,0337	0,0189
5	0,0361	0,1008	0,1563	0,1755	0,1606	0,1277	0,0916	0,0607	0,0378
6	0,0120	0,0504	0,1042	0,1462	0,1606	0,1490	0,1221	0,0911	0,0631
7	0,0034	0,0216	0,0595	0,1045	0,1377	0,1490	0,1396	0,1171	0,0901
8	0,0009	0,0081	0,0298	0,0653	0,1033	0,1304	0,1396	0,1318	0,1126
9	0,0002	0,0027	0,0132	0,0363	0,0689	0,1014	0,1241	0,1318	0,1251
10	0,0000	0,0008	0,0053	0,0181	0,0413	0,0710	0,0993	0,1186	0,1251
11	0,0000	0,0002	0,0019	0,0082	0,0225	0,0452	0,0722	0,0970	0,1137
12	0,0000	0,0001	0,0006	0,0034	0,0113	0,0264	0,0481	0,0728	0,0948
13	0,0000	0,0000	0,0002	0,0013	0,0052	0,0142	0,0296	0,0504	0,0729
14	0,0000	0,0000	0,0001	0,0005	0,0022	0,0071	0,0169	0,0324	0,0521
15	0,0000	0,0000	0,0000	0,0002	0,0009	0,0033	0,0090	0,0194	0,0347
16	0,0000	0,0000	0 0000	0,0000	0,0003	0,0015	0,0045	0,0109	0,0217
17	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0021	0,0058	0,0128
18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0009	0,0029	0,0071
19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0037
20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0006	0,0019
21	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0009
22	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004
23	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002
24	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
25	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Таблица IV

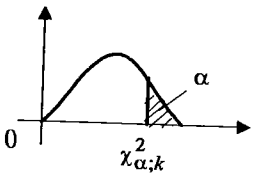
Значения $t_{\gamma,k}$-критерия Стьюдента



Число степеней свободы k	Вероятность γ											
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,95	0,98	0,99
1	0,16	0,32	0,51	0,73	1,00	1,38	1,96	3,08	6,31	12,71	31,82	63,66
2	14	29	44	62	0,82	06	34	1,89	2,92	4,30	6,96	9,92
3	14	28	42	58	76	0,98	25	64	35	3,18	4,54	5,84
4	13	27	41	57	74	94	19	53	13	2,78	3,75	4,60
5	13	27	41	56	73	92	16	48	01	57	36	03
6	0,13	0,26	0,40	0,55	1,72	1,91	1,13	1,44	1,94	2,45	3,14	3,71
7	13	26	40	55	71	90	12	41	89	36	00	50
8	13	26	40	55	70	89	11	40	86	31	2,90	35
9	13	26	40	54	70	88	10	38	83	26	82	25
10	13	26	40	54	70	88	09	37	81	23	76	17
11	0,13	0,26	0,40	0,54	0,70	0,88	1,09	1,36	1,80	2,20	2,72	3,11
12	13	26	39	54	69	87	08	36	78	18	68	05
13	13	26	39	54	69	87	08	35	77	16	65	01
14	13	26	39	54	69	87	08	34	76	14	62	2,98
15	13	26	39	54	69	87	07	34	75	13	60	95
16	0,13	0,26	0,39	0,53	0,69	0,86	1,07	1,34	1,75	2,12	2,58	2,92
17	13	26	39	53	69	86	07	33	74	11	57	90
18	13	26	39	53	69	86	07	33	73	10	55	88
19	13	26	39	53	69	86	07	33	73	09	54	86
20	13	26	39	53	69	86	06	32	72	09	53	84
21	0,13	0,26	0,39	0,53	0,69	0,86	1,06	1,32	1,72	2,08	2,52	2,83
22	13	26	39	53	69	86	06	32	72	07	51	82
23	13	26	39	53	68	86	06	32	71	07	50	81
24	13	26	39	53	68	86	06	32	71	06	49	80
25	13	26	39	53	68	86	06	32	71	06	48	79
26	0,13	0,26	0,39	0,53	0,68	0,86	1,06	1,31	1,71	2,06	2,48	2,78
27	13	26	39	53	68	85	06	31	70	05	47	77
28	13	26	39	53	68	85	06	31	70	05	47	76
29	13	26	39	53	68	85	05	31	70	04	46	76
30	13	26	39	53	68	85	05	31	70	04	46	75
40	0,13	0,25	0,39	0,53	0,68	0,85	1,05	1,30	1,68	2,02	2,42	2,70
60	13	25	39	53	68	85	05	30	67	00	39	66
120	0,13	0,25	0,39	0,53	0,68	0,84	1,04	1,29	1,66	1,98	2,36	2,62
∞	13	25	38	52	67	84	04	28	64	96	33	58

Таблица V

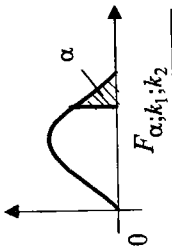
Значения $\chi^2_{\alpha;k}$ критерия Пирсона



Число степеней свободы k	Вероятность α												
	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
1	0,00	0,00	0,00	0,02	0,06	0,15	0,45	1,07	1,64	2,71	3,84	5,41	6,64
2	0,02	0,04	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,60	5,99	7,82	9,21
3	0,11	0,18	0,35	0,58	1,00	1,42	2,37	3,66	4,64	6,25	7,82	9,84	11,3
4	0,30	0,43	0,71	1,06	1,65	2,20	3,36	4,88	5,99	7,78	9,49	11,7	13,3
5	0,55	0,75	1,14	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,1	13,4	15,1
6	0,87	1,13	1,63	2,20	3,07	3,83	5,35	7,23	8,56	10,6	12,6	15,0	16,8
7	1,24	1,56	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,0	14,1	16,6	18,5
8	1,65	2,03	2,73	3,49	4,59	5,53	7,34	9,52	11,0	13,4	15,5	18,2	20,1
9	2,09	2,53	3,32	4,17	5,38	6,39	8,34	10,7	12,2	14,7	16,9	19,7	21,7
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
11	3,05	3,61	4,58	5,58	6,99	8,15	10,3	12,9	14,6	17,3	19,7	22,6	24,7
12	3,57	4,18	5,23	6,30	7,81	9,03	11,3	14,0	15,8	18,5	21,0	24,1	26,2
13	4,11	4,76	5,89	7,04	8,63	9,93	12,3	15,1	17,0	19,8	22,4	25,5	27,7
14	4,66	5,37	6,57	7,79	9,47	10,8	13,3	16,2	18,1	21,1	23,7	26,9	29,1
15	5,23	5,98	7,26	8,55	10,3	11,7	14,3	17,3	19,3	22,3	25,0	28,3	30,6
16	5,81	6,61	7,96	9,31	11,1	12,6	15,3	18,4	20,5	23,5	26,3	29,6	32,0
17	6,41	7,26	8,67	10,1	12,0	13,5	16,3	19,5	21,6	24,8	27,6	31,0	33,4
18	7,02	7,91	9,39	10,9	12,9	14,4	17,3	20,6	22,8	26,0	28,9	32,3	34,8
19	7,63	8,57	10,1	11,6	13,7	15,3	18,3	21,7	23,9	27,2	30,1	33,7	36,2
20	8,26	9,24	10,8	12,4	14,6	16,3	19,3	22,8	25,0	28,4	31,4	35,0	37,6
21	8,90	9,92	11,6	13,2	15,4	17,2	20,3	23,9	26,2	29,6	32,7	36,3	38,9
22	9,54	10,6	12,3	14,0	16,3	18,1	21,3	24,9	27,3	30,8	33,9	37,7	40,3
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
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
25	11,5	12,7	14,6	16,5	18,9	20,9	24,3	28,2	30,7	34,4	37,7	41,7	44,3
26	12,2	13,4	15,4	17,3	19,8	21,8	25,3	29,2	31,8	35,6	38,9	42,9	45,6
27	12,9	14,1	16,1	18,1	20,7	22,7	26,3	30,3	32,9	36,7	40,1	44,1	47,0
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
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
30	14,9	16,3	18,5	20,6	23,4	25,5	29,3	33,5	36,2	40,3	43,8	48,0	50,9

Таблица VI

Значения $F_{\alpha;k_1;k_2}$ -критерия Фишера—Снедекора



α=0,05

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	240	242	244	246	248	249	250	251	252	253	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84

Продолжение табл. VI

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	4,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	0,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	3,60	2,37	2,21	2,10	2,01	1,94	1,83	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00
α=0,01																			
1	4052	4999,5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1.95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

k_1 — число степеней свободы для большой дисперсии, k_2 — для меньшей дисперсии.

Предметный указатель

Учебник

Кремер Наум Шевелевич

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Авторская редакция
Корректор *Г.Б. Костромцова*
Оригинал-макет *Н.В. Спасской*
Оформление художника *В.А. Лебедева*

Лицензия серия ИД № 03562 от 19.12.2000
Подписано в печать 22.08.2003 (с готовых ps-файлов)
Формат 60x88 1/16. Усл. печ. л. 36,0. Уч.-изд. л. 28,5
Тираж 30 000 экз. (2-й завод – 5 000). Заказ № 5508

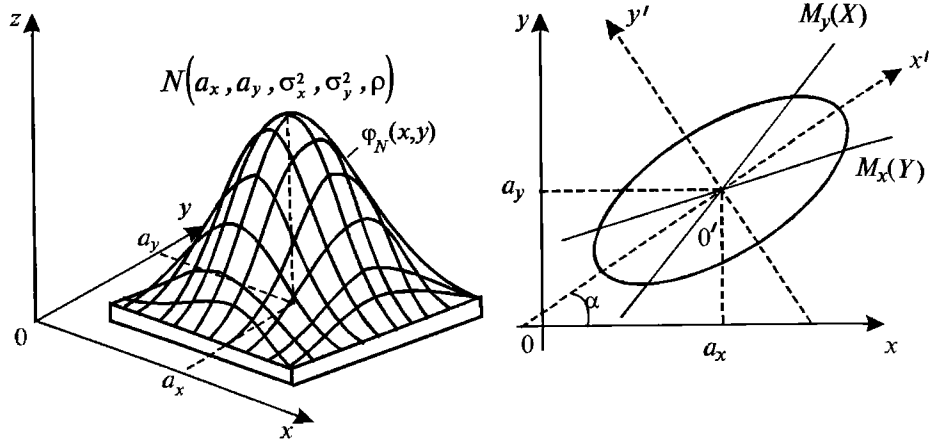
ООО «ИЗДАТЕЛЬСТВО ЮНИТИ-ДАНА»
Генеральный директор *В.Н. Закаидзе*

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

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

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Основная задача — выявление связи между случайными переменными



Коэффициент корреляции

$$\rho = \frac{M(XY) - a_x a_y}{\sigma_x \sigma_y}.$$

Условные математические ожидания и дисперсии

$$M_x(Y) = a_y + \rho\frac{\sigma_y}{\sigma_x}(x - a_x), \quad M_y(X) = a_x + \rho\frac{\sigma_x}{\sigma_y}(y - a_y);$$

$$D_x(Y) = \sigma_y^2(1 - \rho^2), \qquad D_y(X) = \sigma_x^2(1 - \rho^2)$$

Выборочный коэффициент корреляции

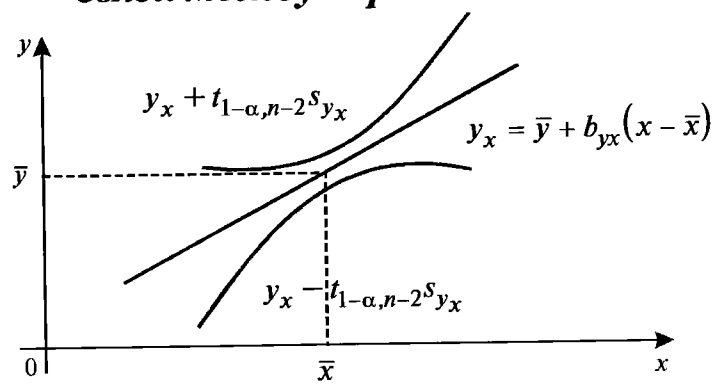
$$r = \frac{\overline{xy} - \overline{x}\,\overline{y}}{s_x s_y}$$

r значим на уровне α, если

$$\frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} > t_{1-\alpha,\ n-2}$$

РЕГРЕССИОННЫЙ АНАЛИЗ

Основная задача — установление формы связи между переменными



Уравнение линейной регрессии

парной	***множественной***
$y_x = \bar{y} + b_1(x - \bar{x}),$	$y_x = X_0'b,$
где	где
$b_1 = \dfrac{\overline{xy} - \bar{x}\,\bar{y}}{s_x^2}$	$b = (X'X)^{-1}X'Y$

Интервальная оценка для $M_x(Y)$

$$y_x - t_{1-\alpha,n-p-1}s_{y_x} \le M_x(Y) \le y_x + t_{1-\alpha,n-p-1}s_{y_x},$$

где $s_{y_x} = s\sqrt{\dfrac{1}{n} + \dfrac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$	где $s_{y_x} = s\sqrt{X_0'(X'X)^{-1}X_0}$

Уравнение регрессии значимо на уровне α, если

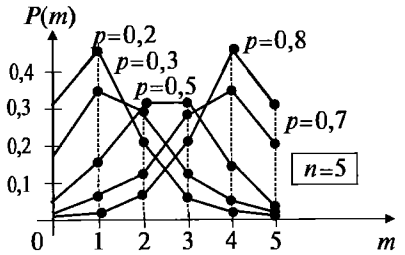
$$F = \frac{R^2(n-p-1)}{(1-R^2)p} > F_{\alpha,p,n-p-1}$$

ОСНОВНЫЕ РАСПРЕДЕЛЕНИЯ И ИХ ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ

$$M(X)=\sum_{i=1}^{n(\infty)} x_i p_i \quad \Longrightarrow \quad D(X)=M\left[X-M(X)\right]^2 \quad \Longleftarrow \quad M(X)=\int_{-\infty}^{+\infty} x\,\varphi(x)\,dx$$

Биномиальное

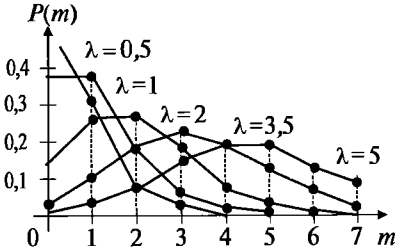
$$P(X=m)=C_n^m p^m q^{n-m}$$



$$M(X)=np, \quad D(X)=npq$$

Пуассона

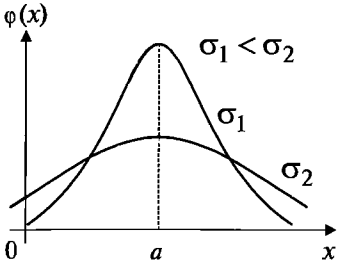
$$P(X=m)=\frac{\lambda^m e^{-\lambda}}{m!}$$



$$M(X)=\lambda, \quad D(X)=\lambda$$

Нормальное

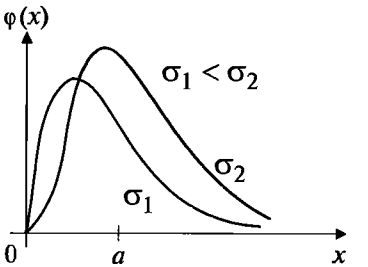
$$\varphi_N(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-a)^2/2\sigma^2}$$



$$M(X)=a, \quad D(X)=\sigma^2$$

Логнормальное

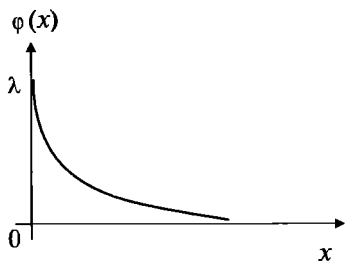
$$\varphi(x)=\frac{1}{\sigma\sqrt{2\pi}x}e^{-(\ln x-\ln a)^2/2\sigma^2}$$



$$M(X)=ae^{\sigma^2/2}, \quad D(X)=a^2e^{\sigma^2}(e^{\sigma^2}-1)$$

Показательное

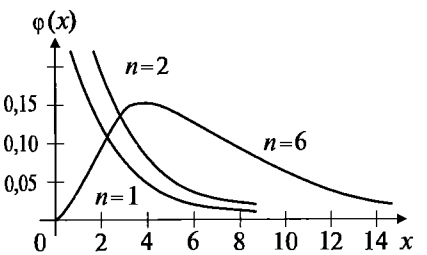
$$\varphi(x)=\lambda e^{-\lambda x} \quad (x>0)$$



$$M(X)=\frac{1}{\lambda}, \quad D(X)=\frac{1}{\lambda^2}$$

Хи-квадрат

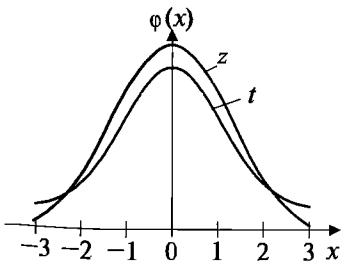
$$\chi^2(n)=\sum_{i=1}^{n} z_i^2, \quad z_i \sim N(0;1)$$



$$M\left[\chi^2(n)\right]=n, \quad D\left[\chi^2(n)\right]=2n$$

Стьюдента

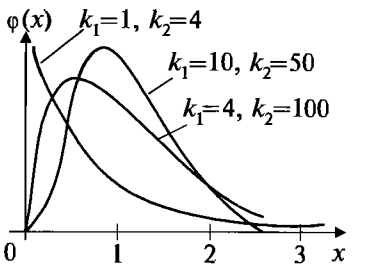
$$t(n)=\frac{z}{\sqrt{\chi^2(n)/n}}, \quad z \sim N(0;1)$$



$$M[t(n)]=0, \quad D[t(n)]=\frac{n}{n-2}$$

Фишера—Снедекора

$$F(m,n)=\frac{n\chi^2(m)}{m\chi^2(n)}$$



$$M[F(m,n)]=\frac{n}{n-2}$$