$\varphi_N(x,y)$

Н.Ш. Кремер

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

УЧЕБНИК
ВТОРОЕ ИЗДАНИЕ

ЮНИТИ
UNITY

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 лег-

че, 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 дру-

гого. 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 н е з н а н и е 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 *элементар-*

---

[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.*

ными 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 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 с в о й с т в а вероятности события.

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

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

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 с и м м е т р и е й 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 ч а с т о будет появляться данное событие в произведенных испытаниях. В этом случае используется с т а т и с т и ч е с к о е о п р е д е л е н и е вероятности.

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

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

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 с л у ч а е в, благоприятствующих событию A, которая определяется непосредственно, без каких-либо испытаний, то $\tilde{P}(A)$ есть доля тех ф а к т и ч е с к и п р о и з в е д е н н ы х и с п ы т а н и й, 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 с в о й с т в а м и.

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 в е р о я т н о с т ь 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.

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

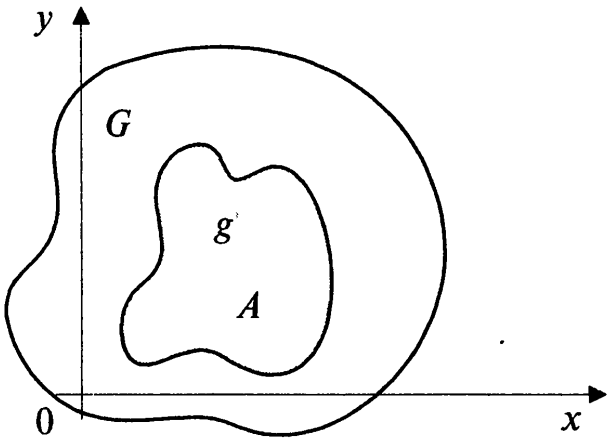


Рис. 1.1

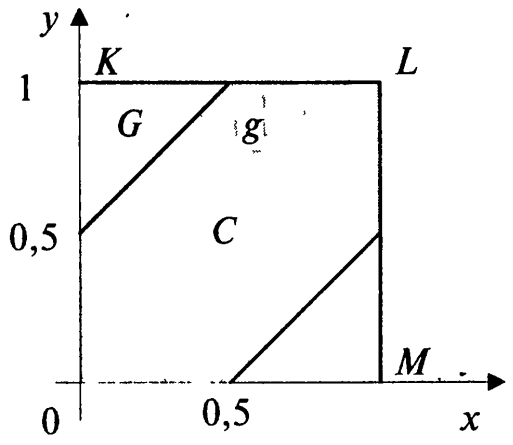


Рис. 1.2

▷ **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**)

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

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. Сколь-

ко 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 $A_1, A_2,\ldots,A_k$ в указанном порядке может быть осуществлен $n_1 n_2 \ldots n_k$ 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 $n_1 n_2 n_3 = 30 \cdot 29 \cdot 28 =$ = 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. $n! = 1 \cdot 2 \ldots n$.

▷ **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

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

а число сочетаний с повторениями из n элементов по m равно

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

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

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

[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**):

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

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 за-

полнения 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 $\bar{A}$, 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

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

и для $n = 365$

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

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.

О п р е д е л е н и е. ***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.

О п р е д е л е н и е. ***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.

О п р е д е л е н и е. ***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) $AC - B$?

Р е ш е н и е. 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

---

[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).

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 $B - A$.

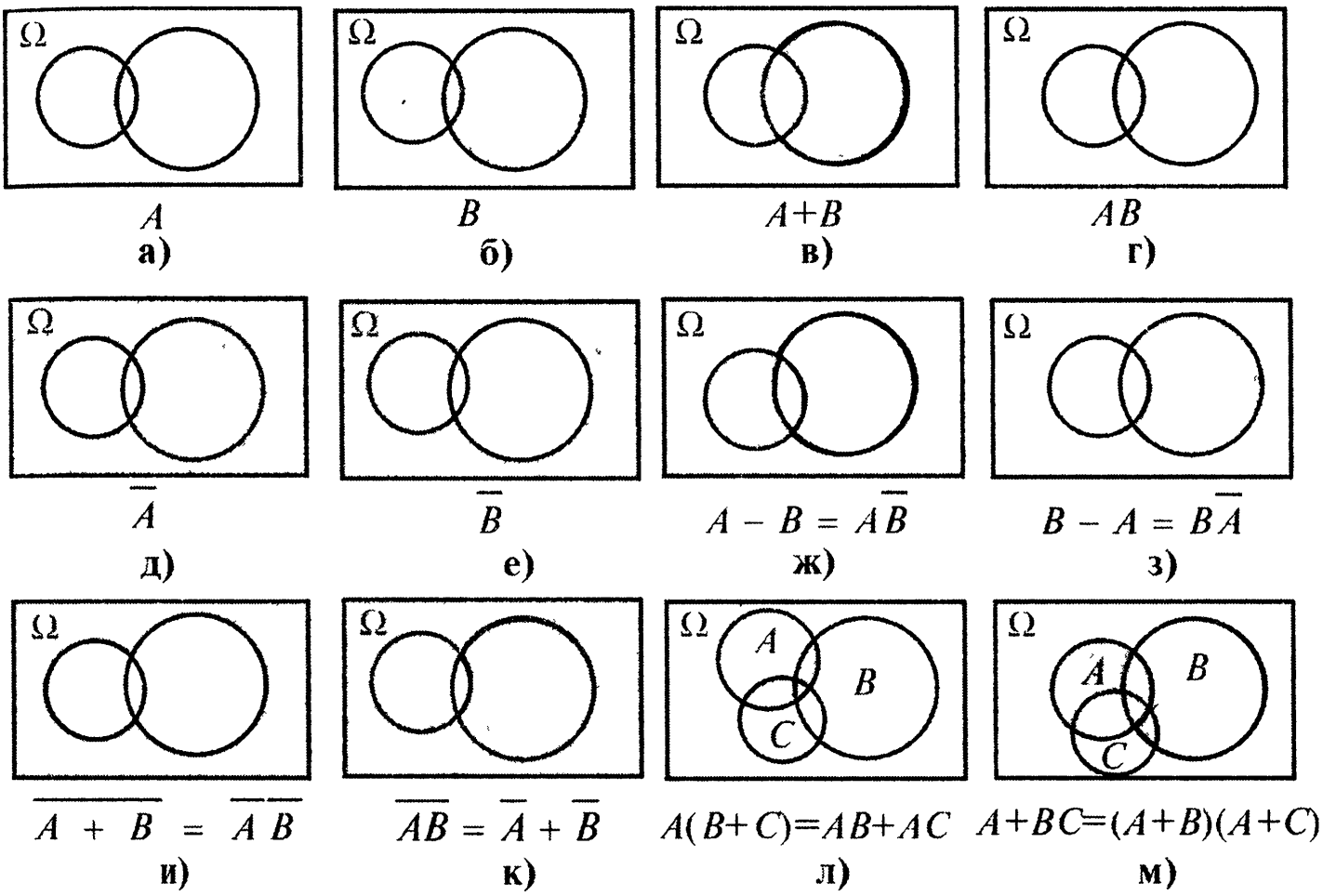


Рис. 1.3

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

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

Р е ш е н и е. 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.

Операции сложения и умножения событий обладают следующими с в о й с т в а м и:

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) с л о ж е н и я 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).

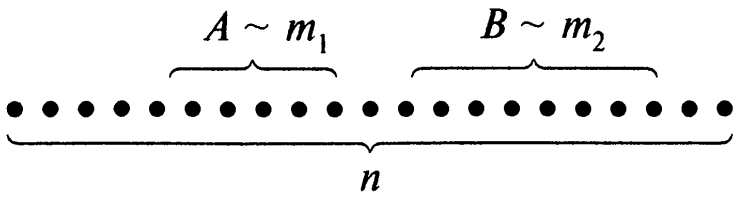


Рис. 1.4

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

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,

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

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

$$P(A) + P(B) + ... + P(K) = 1. \qquad (1.17)$$

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

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

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

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

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

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

$$P(A) + P\left(\overline{A}\right) = 1. \qquad (1.18)$$

□ 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.

Р е ш е н и е. 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 $P(B) = P(C) - P(A) = 0{,}36 - 0{,}13 = 0{,}23$. ◀

З а м е ч а н и е. 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).

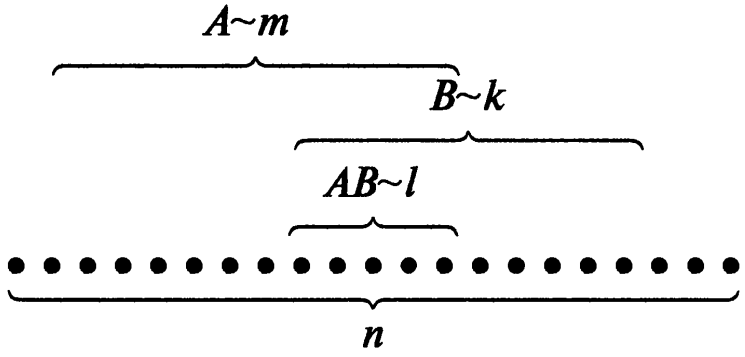


Рис. 1.5

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

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

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 $C = A_1 A_2 A_3 \overline{A}_4$ и по теореме умножения (**1.22**)

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

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

▷**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, *н е з а в и с и м ы.*

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 $P_A(B) \neq P(B) \left(\text{или } P_A(B) \neq P(B)\right)$, 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:

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

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**.

Р е ш е н и е. 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 $P_A(B)\neq$ $\neq P_{\bar{A}}(B)\left(\frac{2}{4}\neq\frac{3}{4}\right)$, 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.

Д л я н е з а в и с и м ы х 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?

Р е ш е н и е. 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 о п р е д е л е н и я независимости д в у х событий. Для определения независимости н е с к о л ь к и х 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

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

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.*

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

□ 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

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

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**):

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

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

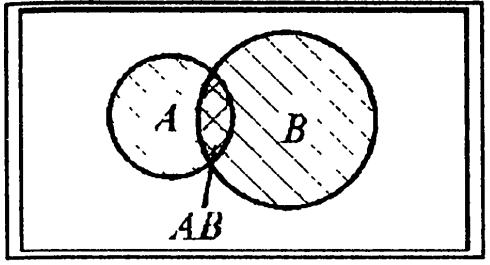


Рис. 1.6

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:

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

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

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

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 — н е з а в и с и м ы е, то

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

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

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

(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 — $\overline{F} = \overline{A_1 + A_2 + A_3} = \overline{A}_1\overline{A}_2\overline{A}_3$, 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 $P(B) = P(\overline{A})P(\overline{A})P(A) = 0{,}4\cdot0{,}4\cdot0{,}6 =$ $= 0{,}096$.

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,

$$P(C) = P(A) + P(A)P(A) + P(A)P(A)\cdot P(A) =$$
$$= 0{,}6 + 0{,}4\cdot0{,}6 + 0{,}4\cdot0{,}4\cdot0{,}6 = 0{,}936\,.$$ ▶

▷**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 $A_1, A_2, \ldots, A_i, \ldots, A_n$ определится по формуле (**1.29**):

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

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

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

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 ($i=1, 2, \ldots$);

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:

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

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

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

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

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

$$=0{,}7\cdot0{,}8+0{,}7\cdot0{,}2\cdot0{,}8+0{,}3\cdot0{,}8\cdot0{,}7+0{,}3\cdot0{,}2\cdot0{,}7\cdot0{,}8=0{,}8736.$$ ▶

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:

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

□ 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

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

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

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

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 $A_1,A_2,\ldots,A_n$, образующих полную группу событий, произошло и необходимо произвести количественную переоценку *априорных* вероятностей этих гипотез $P(A_1)$, $P(A_2),\ldots,P(A_n)$, известных до испытания, т.е. надо найти *апостериорные* (получаемые после проведения испытания) условные вероятности гипотез $P_F(A_1),P_F(A_2),\ldots,P_F(A_n)$.

□ 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

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

или с учетом **(1.31)**

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

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:

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

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

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

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

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 $A_i A_j = \varnothing (i \neq j)$.

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. $0\le P(A)\le 1$.

4. $P(A)\le P(B)$, если $A\subset B$.

5. $P(A+B)=P(A)+P(B)-P(AB)$.

6. $P(A+B)\le P(A)+P(B)$.

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 $A_1, A_2, \ldots, A_n, \ldots$, ($A_iA_j=\varnothing$ при $i\ne j$), 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 $P(A)\ne 0$.

---

[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 хо-

тя 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*

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

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

$$B_2 = A_1 A_2 A_3 + A_1 A_2 A_3 + A_1 A_2 A_3,$$

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

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

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 $p(A_i)=p$, $p(\overline{A}_i)=q$, $i = 1, 2, \ldots, n$. В связи с тем, что комбинации между собой несовместны, по теореме сложения вероятностей получим

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

▷ **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**):

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

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

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

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

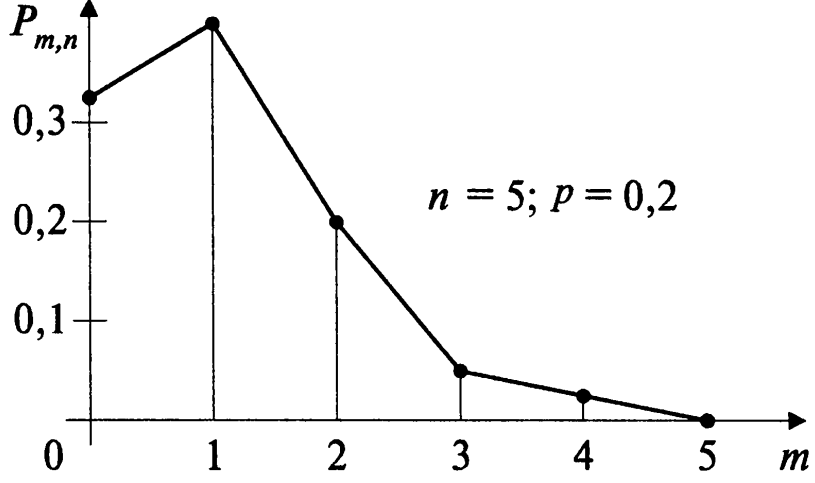


Рис. 2.1

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)** $5 \cdot 0{,}2 - 0{,}8 \le m_0 \le 5 \cdot 0{,}2 + 0{,}2$ или $0{,}2 \le m_0 \le 1{,}2$. 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)** $n \cdot \frac{1}{6} - \frac{5}{6} \le 10 \le n \cdot \frac{1}{6} + \frac{1}{6}$ или $n - 5 \le 60 \le n + 1$, откуда $59 \le n \le 65$, 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 $\lim\limits_{n\to\infty}\left(1-\frac{1}{n}\right) = \lim\limits_{n\to\infty}\left(1-\frac{2}{n}\right) = \ldots = \lim\limits_{n\to\infty}\left(1-\frac{m-1}{n}\right) = 1,$

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

$$\lim_{n\to\infty} P_{m,n} = \frac{\lambda^m}{m!} e^{-\lambda}. \blacksquare$$

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 $p = \text{const}$. 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**):

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

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*;

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

(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 свойства функции $\Phi(x)$.

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

□ $\Phi(-x) = \frac{2}{\sqrt{2\pi}} \int_0^{-x} e^{-t^2/2} dt$. 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. $\Phi'(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}$, и всегда положительна, то $\Phi(x)$ монотонно возрастает на всей числовой прямой.

Найдем

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

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.

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

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)**

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

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

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

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 прак-

тически 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.

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

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

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

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

▷ **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 $npq = 10\,000 \cdot 0{,}5(1 - 0{,}5) = 250 \ge 20$ 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 стра-

ховую 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 $npq = 10\,000 \cdot 0{,}005 \cdot 0{,}995 = 49{,}75 \geq 20$ выполнено).

По условию задачи

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

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. Полиноминальная 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 $\overline{A}$) 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

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

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

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

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)?

2.33. Вероятность того, что дилер, торгующий ценными бумагами, продаст их, равна 0,7. Сколько должно быть ценных бумаг, чтобы можно было утверждать с вероятностью 0,996, что доля проданных среди них отклонится от 0,7 не более, чем на 0,04 (по абсолютной величине)?

2.34. У страховой компании имеются 10 000 клиентов. Каждый из них, страхуясь от несчастного случая, вносит 500 руб. Вероятность несчастного случая 0,0055, а страховая сумма, выплачиваемая пострадавшему, составляет 50 000 руб. Какова вероятность того, что: а) страховая компания потерпит убыток; б) на выплату страховых сумм уйдет более половины всех средств, поступивших от клиентов?

2.35. Первый прибор состоит из 10 узлов, второй из 8 узлов. За время t каждый из узлов первого прибора выходит из строя, независимо от других, с вероятностью 0,1, второго — с вероятностью 0,2. Найти вероятность того, что за время t в первом приборе выйдет из строя хотя бы один узел, а во втором — по крайней мере два узла.

2.36. Студент рассматриваемого вуза по уровню подготовленности с вероятностью 0,3 является «слабым», с вероятностью 0,5 — «средним», с вероятностью 0,2 — «сильным». Какова вероятность того, что из наудачу выбранных 6 студентов вуза: а) число «слабых», «средних» и «сильных» окажется одинаковым; б) число «слабых» и «сильных» окажется одинаковым?

Глава 3 Случайные величины

3.1. Понятие случайной величины. Закон распределения дискретной случайной величины

Одним из важнейших понятий теории вероятностей является понятие случайной величины.

Под *случайной величиной* понимается переменная, которая в результате испытания в зависимости от случая принимает одно из возможного множества своих значений (какое именно — заранее не известно).

Примеры случайных величин:

1) число родившихся детей в течение суток в г. Москве;
2) количество бракованных изделий в данной партии;
3) число произведенных выстрелов до первого попадания;
4) дальность полета артиллерийского снаряда;
5) расход электроэнергии на предприятии за месяц.

Случайная величина называется *дискретной* (*прерывной*), если множество ее значений конечное, или бесконечное, но счетное[1].

Под *непрерывной* случайной величиной будем понимать величину, бесконечное несчетное множество значений которой есть некоторый интервал (конечный или бесконечный) числовой оси[2].

Так, в приведенных выше примерах 1—3 имеем дискретные случайные величины (в примерах 1 и 2 — с конечным множеством значений; в примере 3 — с бесконечным, но счетным множеством значений); а в примерах 4 и 5 — непрерывные случайные величины.

Теоретико-множественная трактовка основных понятий теории вероятностей позволяет дать следующее определение случайной величины.

Определение. ***Случайной величиной*** *X называется функция, заданная на множестве элементарных исходов (или в пространстве элементарных событий)*[3], т.е.

---

[1] См. сноску на с. 59.

[2] Строгое определение непрерывной случайной величины дано ниже.

[3] В случае бесконечного несчетного множества элементарных событий Ω это определение нуждается в уточнении (связанном с измеримостью функции $f(\omega)$ относительно σ-алгебры S), которое здесь не приводится, так как выходит за рамки данной книги.

$$X = f(\omega),$$

где ω — элементарный исход (или элементарное событие, принадлежащее пространству Ω, т.е. $\omega \in \Omega$).

Для дискретной случайной величины множество Ξ возможных значений случайной величины, т.е. функции $f(\omega)$, конечно или счетно, для непрерывной — бесконечно и несчетно.

Убедимся, например, в том, что случайная величина X – число дней во взятом наудачу месяце года (невисокосного) есть функция элементарных исходов (событий) ω, т.е. $X = f(\omega)$. В результате испытания — розыгрыша (выбора наудачу) месяца года – все множество элементарных исходов (пространство элементарных событий) Ω может быть представлено в виде

$$\Omega = \{\omega_1, \omega_2, \omega_3, ..., \omega_{12}\},$$

где ω_1, ω_2, ω_3, ..., ω_{12} — соответственно 1-й, 2-й, 3-й, ..., 12-й месяц года.

Так как $X(\omega_1)=31$, $X(\omega_2)=28$, $X(\omega_3)=31$, $X(\omega_4)=30,...,X(\omega_{12})=31$, то число дней во взятом наудачу месяце года (случайная величина X) есть функция элементарных исходов (событий) ω.

Случайные величины будем обозначать прописными буквами латинского алфавита X, Y, Z,..., а их значения — соответствующими строчными буквами x, y, z,... .

Наиболее полным, исчерпывающим описанием случайной величины является ее закон распределения.

О п р е д е л е н и е. ***Законом распределения*** *случайной величины называется всякое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями.*

Про случайную величину говорят, что она «распределена» по данному закону распределения или «подчинена» этому закону распределения.

Для д и с к р е т н о й случайной величины закон распределения может быть задан в виде таблицы, аналитически (в виде формулы) и графически.

Простейшей формой задания закона распределения дискретной случайной величины X является таблица (матрица), в которой перечислены в порядке возрастания все возможные значения случайной величины и соответствующие их вероятности, т.е.

X:	x_1	x_2	...	x_i	...	x_n
	p_1	p_2	...	p_i	...	p_n

или $$X = \begin{pmatrix} x_1 & x_2 & \dots x_n \\ p_1 & p_2 & \dots p_n \end{pmatrix}.$$

Такая таблица называется *рядом распределения* дискретной случайной величины.

События $X=x_1$, $X=x_2$,..., $X=x_n$, состоящие в том, что в результате испытания случайная величина X примет соответственно значения x_1, x_2, ..., x_n, являются несовместными и единственно возможными (ибо в таблице перечислены все возможные значения случайной величины), т.е. образуют полную группу. Следовательно, сумма их вероятностей равна 1. Таким образом, *для любой дискретной случайной величины*

$$\sum_{i=1}^{n} P(X = x_i) = \sum_{i=1}^{n} p_i = 1. \qquad \textbf{(3.1)}$$

(Эта единица как-то р а с п р е д е л е н а между значениями случайной величины, отсюда и термин «распределение».)

Ряд распределения может быть изображен графически, если по оси абсцисс откладывать значения случайной величины, а по оси ординат — соответствующие их вероятности. Соединение полученных точек образует ломаную, называемую *многоугольником* или *полигоном распределения вероятностей* (рис. 3.1).

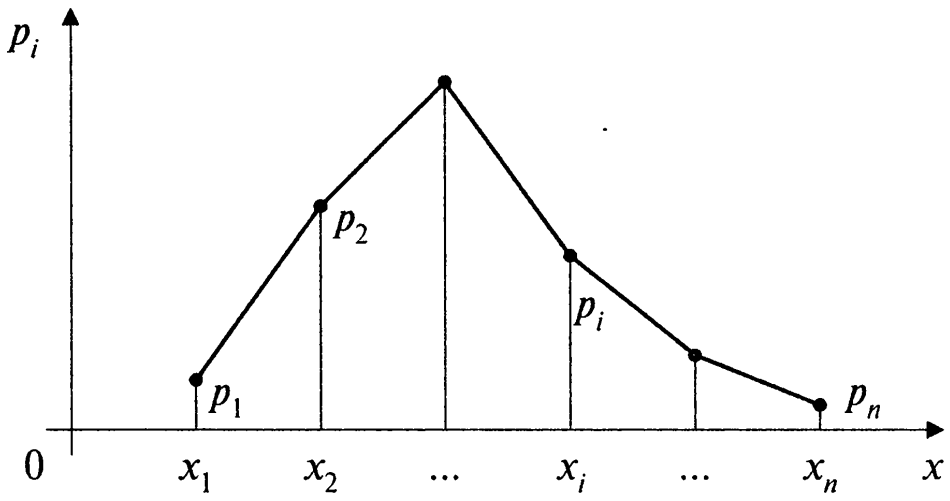


Рис. 3.1

▷ **Пример 3.1.** В лотерее разыгрываются: автомобиль стоимостью 5000 ден. ед., 4 телевизора стоимостью 250 ден. ед., 5 видеомагнитофонов стоимостью 200 ден. ед. Всего продается 1000 билетов по 7 ден. ед. Составить закон распределения чистого выигрыша, полученного участником лотереи, купившим один билет.

Р е ш е н и е. Возможные значения случайной величины X — чистого выигрыша на один билет — равны $0-7=-7$ ден. ед. (если билет не выиграл), $200-7=193$, $250-7=243$, $5000-7=4993$ ден. ед. (если на билет выпал выигрыш соответственно видеомагнитофона, телевизора или автомобиля). Учитывая, что из 1000 билетов число невыигравших составляет 990, а указанных выигрышей соответственно 5, 4 и 1, и используя классическое определение вероятности, получим:

$$P(X=-7)=990/1000=0{,}990;\ P(X=193)=5/1000=0{,}005;$$

$$P(X=243)=4/1000=0{,}004;\ P(X=4993)=1/1000=0{,}001,$$

т.е. ряд распределения

X:	x_i	-7	193	243	4993
	p_i	0,990	0,005	0,004	0,001

▶

▷ **Пример 3.2.** Вероятности того, что студент сдаст семестровый экзамен в сессию по дисциплинам A и $Б$, равны соответственно 0,7 и 0,9. Составить закон распределения числа семестровых экзаменов, которые сдаст студент.

Р е ш е н и е. Возможные значения случайной величины X — числа сданных экзаменов — 0, 1, 2.

Пусть A_i — событие, состоящее в том, что студент сдаст i-й экзамен ($i=1, 2$). Тогда вероятности того, что студент сдаст в сессию 0, 1, 2 экзамена, будут соответственно равны (считаем события A_1 и A_2 независимыми):

$$P(X=0)=P\left(\overline{A}_1\overline{A}_2\right)=P\left(\overline{A}_1\right)P\left(\overline{A}_2\right)=$$

$$=(1-0{,}7)(1-0{,}9)=0{,}3\cdot 0{,}1=0{,}03;$$

$$P(X=1)=P\left(A_1\overline{A}_2+\overline{A}_1A_2\right)=P\left(A_1\right)P\left(\overline{A}_2\right)+P\left(\overline{A}_1\right)P\left(A_2\right)=$$

$$=0{,}7\cdot 0{,}1+0{,}3\cdot 0{,}9=0{,}34,$$

$$P(X=2)=P\left(A_1A_2\right)=P\left(A_1\right)P\left(A_2\right)=0{,}7\cdot 0{,}9=0{,}63.$$

Итак, ряд распределения случайной величины

X:

x_i	0	1	2
p_i	0,03	0,34	0,63

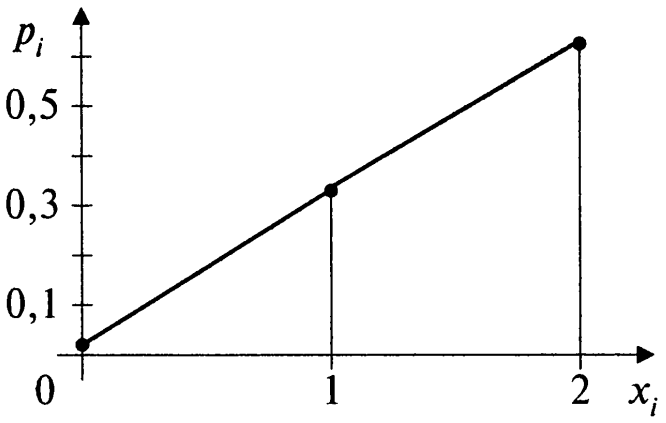


Рис. 3.2

На рис. 3.2 полученный ряд распределения представлен графически в виде многоугольника (полигона) распределения вероятностей. ▶

3.2. Математические операции над случайными величинами

Вначале введем понятие независимости случайных величин.

Две случайные величины называются *независимыми*, если закон распределения одной из них не меняется от того, какие возможные значения приняла другая величина. Так, если дискретная случайная величина X может принимать значения x_i ($i = 1, 2, ..., n$), а случайная величина Y — значения y_j ($j = 1, 2, ..., m$), то независимость дискретных случайных величин X и Y означает независимость событий $X = x_i$ и $Y = y_j$ при любых $i = 1, 2, ..., n$ и $j = 1, 2, ..., m$. В противном случае случайные величины называются *зависимыми*.

Например, если имеются билеты двух различных денежных лотерей, то случайные величины X и Y, выражающие соответственно выигрыш по каждому билету (в денежных единицах), будут независимыми, так как при любом выигрыше по билету одной лотереи (например, при $X = x_i$) закон распределения выигрыша по другому билету (Y) не изменится. Если же случайные величины X и Y выражают выигрыш по билетам одной денежной лотереи, то в этом случае X и Y являются зависимыми, ибо любой выигрыш по одному билету ($X = x_i$)

приводит к изменению вероятностей выигрыша по другому билету (Y), т.е. к изменению закона распределения Y.

В дальнейшем понятие независимости случайных величин будет уточнено (см. § 5.5).

Определим математические операции над дискретными случайными величинами.

Пусть даны две случайные величины:

X:

x_i	x_1	x_2	...	x_n
p_i	p_1	p_2	...	p_n

,

Y:

y_j	y_1	y_2	...	y_m
p_j	p_1	p_2	...	p_m

Произведением kX случайной величины X на постоянную величину k называется случайная величина, которая принимает значения kx_i с теми же вероятностями p_i ($i = 1, 2, ..., n$).

m-й степенью случайной величины X, т.е. X^m, называется случайная величина, которая принимает значения x_i^m с теми же вероятностями p_i ($i = 1, 2, ..., n$).

▷ **Пример 3.3.** Дана случайная величина

X:

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

Найти закон распределения случайных величин: а) $Y = 3X$; б) $Z = X^2$.

Р е ш е н и е. а) Значения случайной величины Y будут: $3(-2) = -6$; $3 \cdot 1 = 3$; $3 \cdot 2 = 6$ с теми же вероятностями 0,5; 0,3; 0,2, т.е.

Y:

y_i	-6	3	6
p_i	0,5	0,3	0,2

б) Значения случайной величины Z будут: $(-2)^2 = 4$, $1^2 = 1$, $2^2 = 4$ с теми же вероятностями 0,5; 0,3; 0,2. Так как значение $Z = 4$ может быть получено возведением в квадрат значений (-2) с вероятностью 0,5 и $(+2)$ с вероятностью 0,2, то по теореме сложения $P(Z = 4) = 0,5 + 0,2 = 0,7$. Итак, закон распределения случайной величины

Z:

z_i	1	4
p_i	0,3	0,7

▶

Суммой (*разностью* или *произведением*) *случайных величин* X *и* Y называется случайная величина, которая принимает все возможные значения вида x_i+y_j (x_i-y_j или $x_i \cdot y_j$), где $i=1,2,...,n$; $j=1,2,...,m$, с вероятностями p_{ij} того, что случайная величина X примет значение x_i, а Y — значение y_j:

$$p_{ij} = P\left[\left(X = x_i\right)\left(Y = y_j\right)\right].$$

Если случайные величины X и Y независимы, т.е. независимы любые события $X=x_i$, $Y=y_j$, то по теореме умножения вероятностей для независимых событий

$$p_{ij} = P\left(X = x_i\right) \cdot P\left(Y = y_j\right) = p_i \cdot p_j. \qquad \textbf{(3.2)}$$

Замечание. Приведенные выше определения операций над дискретными случайными величинами нуждаются в уточнении: так как в ряде случаев одни и те же значения x_i^m, $x_i \pm y_j$, $x_i y_j$ могут получаться разными способами при различных x_i, y_j с вероятностями p_i, p_{ij}, то вероятности таких повторяющихся значений находятся сложением полученных вероятностей p_i или p_{ij} (см. примеры **3.3***б* и **3.4**).

▷ **Пример 3.4.** Даны законы распределения двух независимых случайных величин:

X:

x_i	0	2	4
p_i	0,5	0,2	0,3

Y:

y_j	−2	0	2
p_j	0,1	0,6	0,2

Найти закон распределения случайных величин: а) $Z=X-Y$; б) $U=XY$.

Р е ш е н и е. а) Для удобства нахождения всех значений разности $Z=X-Y$ и их вероятностей составим вспомогательную таблицу, в каждой клетке которой поместим в левом углу значения разности $Z=X-Y$, а в правом углу — вероятности этих значений, полученные в результате перемножения вероятностей соответствующих значений случайных величин X и Y.

x_i \ y_j	p_j	−2	0	2
p_i		0,1	0,6	0,3
0	0,5	2 / 0,05	0 / 0,30	−2 / 0,15
2	0,2	4 / 0,02	2 / 0,12	0 / 0,06
4	0,3	6 / 0,03	4 / 0,18	2 / 0,09

Например, если $X = 4$ (последняя строка таблицы), а $Y = -2$ (третий столбец таблицы), то случайная величина $Z=X-Y$ принимает значение $Z = 4 - (-2) = 6$ с вероятностью $P(Z = 6) = P(X=4)P(Y=-2)=0{,}3 \cdot 0{,}1=0{,}03$ (эти числа $Z=6$ и $P=0{,}03$ находятся в клетке на пересечении последней строки и третьего столбца).

Так как среди 9 значений Z имеются повторяющиеся, то соответствующие вероятности их складываем по теореме сложения вероятностей (см. замечание на с. 95). Например, значение $Z = X - Y = 2$ может быть получено, когда $X=0$, $Y=-2$ (с вероятностью 0,05); $X=2$, $Y=0$ (с вероятностью 0,12); $X=4$; $Y=2$ (с вероятностью 0,09), поэтому

$$P(Z = 2) = 0{,}15 + 0{,}12 + 0{,}09 = 0{,}26 \text{ и т.д.}$$

В результате получим распределение

Z:

z_k	−2	0	2	4	6
p_k	0,15	0,36	0,26	0,20	0,03

Убеждаемся в том, что условие $\sum_{i=1}^{5} p_i = 1$ выполнено.

б) Распределение $U=XY$ находится аналогично п. а).

U:

u_k	−8	−4	0	4	8
p_k	0,03	0,02	0,80	0,06	0,09

3.3. Математическое ожидание дискретной случайной величины

Закон (ряд) распределения дискретной случайной величины дает исчерпывающую информацию о ней, так как позволяет вычислить вероятности любых событий, связанных со случайной величиной. Однако такой закон (ряд) распределения бывает трудно обозримым, не всегда удобным (и даже необходимым) для анализа. Рассмотрим, например, задачу.

Задача. Известны законы распределения случайных величин X и Y — числа очков, выбиваемых 1-м и 2-м стрелками.

X:

x_i	0	1	2	3	4	5	6	7	8	9	10
p_i	0,15	0,11	0,04	0,05	0,04	0,10	0,10	0,04	0,05	0,12	0,20

Y:

y_j	0	1	2	3	4	5	6	7	8	9	10
p_j	0,01	0,03	0,05	0,09	0,11	0,24	0,21	0,10	0,10	0,04	0,02

Необходимо выяснить, какой из двух стрелков стреляет лучше.

Рассматривая ряды распределения случайных величин X и Y, ответить на этот вопрос далеко не просто из-за обилия числовых значений. К тому же у первого стрелка достаточно большие вероятности (например, больше 0,1) имеют крайние значения числа выбиваемых очков ($X = 0; 1$ и $X = 9; 10$), а у второго стрелка — промежуточные значения ($Y = 4; 5; 6$) (см. многоугольники распределения вероятностей X и Y на рис. 3.3).

Очевидно, что из двух стрелков лучше стреляет тот, кто в с р е д н е м выбивает большее количество очков.

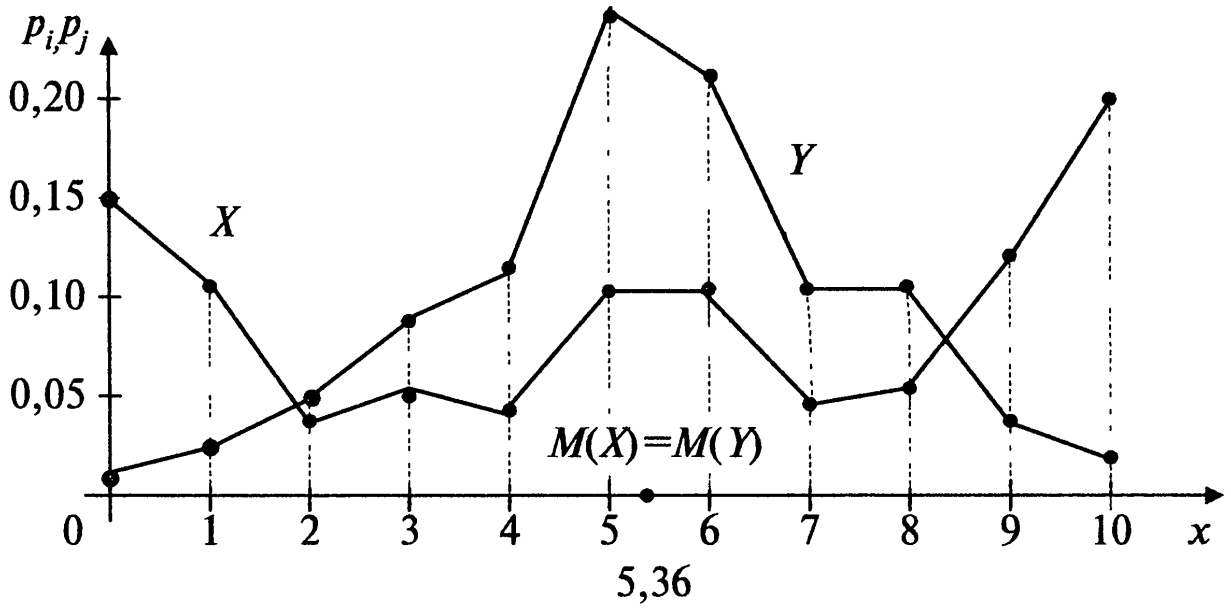


Рис. 3.3

Таким средним значением случайной величины является ее математическое ожидание[1].

О п р е д е л е н и е. ***Математическим ожиданием, или средним значением, $M(X)$ дискретной случайной величины X называется сумма произведений всех ее значений на соответствующие им вероятности***[2]:

$$M(X)=\sum_{i=1}^{n} x_i p_i. \qquad \textbf{(3.3)}$$

▷**Пример 3.5.** Вычислить $M(X)$ и $M(Y)$ в задаче о стрелках.

Р е ш е н и е. По формуле **(3.3)**

$$M(X)=0\cdot 0{,}15+1\cdot 0{,}10+2\cdot 0{,}04+\ldots+9\cdot 0{,}12+10\cdot 0{,}20=5{,}36,$$

$$M(Y)=0\cdot 0{,}01+1\cdot 0{,}03+2\cdot 0{,}05+\ldots+9\cdot 0{,}04+10\cdot 0{,}02=5{,}36,$$

т.е. среднее число выбиваемых очков у двух стрелков одинаковое. ▶

▷ **Пример 3.6.** Вычислить $M(X)$ для случайной величины X — чистого выигрыша по данным примера **3.1**.

Р е ш е н и е. По формуле **(3.3)**

$$M(X)=(-7)\cdot 0{,}990+193\cdot 0{,}005+243\cdot 0{,}004+4993\cdot 0{,}001=0,$$

т.е. средний выигрыш равен нулю. Полученный результат означает, что вся выручка от продажи билетов лотереи идет на выигрыши. ▶

Обратим внимание на м е х а н и ч е с к у ю и н т е р п р е т а ц и ю математического ожидания. Если предположить, что каждая материальная точка с абсциссой x_i имеет массу, равную p_i ($i = 1, 2, \ldots, n$), а вся е д и н и ч н а я масса

---

[1] Происхождение термина «математическое ожидание» связано с начальным периодом возникновения теории вероятностей, когда область ее применения ограничивалась азартными играми. Игрока интересовало среднее значение ожидаемого выигрыша или, иначе, математическое ожидание выигрыша.

[2] Для математического ожидания случайной величины X в литературе также используются обозначения $E(X)$, $\overline{X}$.

$\left(\sum_{i=1}^{n} p_i = 1\right)$ распределена между этими точками, то математическое ожидание представляет собой абсциссу центра масс системы материальных точек. Так, для систем материальных точек, соответствующим распределениям X и Y в примере 3.5, центры масс совпадают: $M(X) = M(Y) = 5{,}36$ (см. рис. 3.3).

*Если дискретная случайная величина X принимает бесконечное, но счетное множество значений $x_1, x_2, \ldots, x_n, \ldots$, то **математическим ожиданием**, или **средним значением**, такой дискретной случайной величины называется сумма ряда* (если он абсолютно сходится):

$$M(X) = \sum_{i=1}^{\infty} x_i p_i. \qquad \textbf{(3.4)}$$

Так как ряд **(3.4)** может и расходиться, то соответствующая случайная величина может и не иметь математического ожидания. Например, случайная величина X с рядом распределения

X:							
	x_i	2	2^2	2^3	...	2^i	...
	p_i	1/2	$1/2^2$	$1/2^3$	...	$1/2^i$	...

не имеет математического ожидания, ибо сумма ряда $\sum_{i=1}^{\infty} 2^i/2^i = \sum_{i=1}^{\infty} 1$ равна ∞. На практике, как правило, множество возможных значений случайной величины распространяется лишь на ограниченный участок оси абсцисс и, значит, математическое ожидание существует.

Рассмотрим свойства математического ожидания.

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

$$M(C) = C. \qquad \textbf{(3.5)}$$

□ Постоянную величину C можно рассматривать как величину, принимающую значение C с вероятностью 1. Поэтому $M(C) = C \cdot 1 = 1$. ■

2. *Постоянный множитель можно выносить за знак математического ожидания*, т.е.

$$M(kX) = kM(X). \qquad \textbf{(3.6)}$$

Так как случайная величина kX принимает значения kx_i ($i = 1, 2,..., n$), то $M(kX) = \sum_{i=1}^{n}(kx_i)p_i = k\sum_{i=1}^{n} x_i p_i = kM(X)$. ■

3. *Математическое ожидание алгебраической суммы конечного числа случайных величин равно такой же сумме их математических ожиданий*, т.е.[1]

$$M(X \pm Y) = M(X) \pm M(Y). \qquad (3.7)$$

□ В соответствии с определением суммы и разности случайных величин (см. § 3.2) $X+Y$ ($X-Y$) представляют случайную величину, которая принимает значения x_i+y_j (x_i-y_j) ($i = 1, 2,..., n$; $j = 1, 2,..., m$) с вероятностями $p_{ij} = P[(X = x_i)(Y = y_j)]$.

Поэтому

$$M(X \pm Y) = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i \pm y_j)p_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i p_{ij} \pm \sum_{i=1}^{n}\sum_{j=1}^{m} y_j p_{ij}.$$

Так как в первой двойной сумме x_i не зависит от индекса j, по которому ведется суммирование во второй сумме, и аналогично во второй двойной сумме y_j не зависит от индекса i, то

$$M(X \pm Y) = \sum_{i=1}^{n} x_i \sum_{j=1}^{m} p_{ij} \pm \sum_{j=1}^{m} y_j \sum_{i=1}^{n} p_{ij} = \sum_{i=1}^{n} x_i p_i \pm \sum_{j=1}^{m} y_j p_j =$$

$$= M(X) \pm M(Y). \blacksquare$$

4. *Математическое ожидание произведения конечного числа независимых случайных величин равно произведению их математических ожиданий*[2]:

$$M(XY) = M(X)M(Y).$$

□ В соответствии с определением произведения случайных величин (см. § 3.2), XY представляет собой случайную величину, которая принимает значения x_iy_j ($i = 1, 2,..., n$; $j = 1, 2,..., m$) с вероятностями $p_{ij} = P[(X = x_i)\ (Y = y_j)]$, причем в силу независимости X и Y $p_{ij} = p_i p_j$. Поэтому

---

[1] Записываем свойство для двух случайных величин.

[2] Записываем свойство для двух случайных величин; случай зависимых случайных величин рассматривается в § 5.6 — см. (**5.40**).

$$M(XY)=\sum_{i=1}^{n}\sum_{j=1}^{m}x_i y_j p_{ij}=\sum_{i=1}^{n}\sum_{j=1}^{m}x_i y_j p_i p_j=\sum_{i=1}^{n}x_i p_i\cdot\sum_{j=1}^{m}y_j p_j=$$

$$=M(X)\cdot M(Y).\ \blacksquare$$

5. *Если все значения случайной величины увеличить (уменьшить) на постоянную C, то на эту же постоянную C увеличится (уменьшится) математическое ожидание этой случайной величины:*

$$M(X\pm C)=M(X)\pm C. \qquad \textbf{(3.8)}$$

□ Учитывая свойства **3** и **1** математического ожидания, получим

$$M(X\pm C)=M(X)\pm M(C)=M(X)\pm C.\ \blacksquare$$

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

$$M[X-M(X)]=0. \qquad \textbf{(3.9)}$$

□ Пусть постоянная C есть математическое ожидание[1] $a=M(X)$, т.е. $C=a$. Тогда, используя свойство **5**, получим

$$M(X-a)=M(X)-a=a-a=0.\ \blacksquare$$

▷ **Пример 3.7.** Найти математическое ожидание случайной величины $Z=8X-5Y+7$, если известно, что $M(X)=3$, $M(Y)=2$.

Р е ш е н и е. Используя свойства **1**, **2**, **3** математического ожидания, найдем

$$M(Z)=8M(X)-5M(Y)+7=8\cdot 3-5\cdot 2+7=21.\ \blacktriangleright$$

3.4. Дисперсия дискретной случайной величины

Только математическое ожидание не может в достаточной степени характеризовать случайную величину.

В задаче о стрелках (см.§ 3.3) мы убедились в том, что $M(X)=M(Y)=5{,}36$, т.е. среднее количество выбиваемых очков у двух

[1] См. замечание на с. 104.

стрелков одинаковое. Но если у 1-го стрелка, как отмечено выше, значительные вероятности имеют крайние значения, сильно отличающиеся от среднего $M(X)$, то у 2-го, наоборот, — значения, близкие к среднему $M(Y)$ (см. рис. 3.3). Очевидно, лучше стреляет тот стрелок, у которого при равенстве средних значений числа выбитых очков меньше о т к л о н е н и я (разброс, вариация, рассеяние) этого числа относительно среднего значения.

В качестве такой характеристики рассматривается д и с п е р с и я случайной величины. Слово дисперсия означает «рассеяние».

О п р е д е л е н и е. ***Дисперсией*** $D(X)$ *случайной величины* X *называется математическое ожидание квадрата ее отклонения от математического ожидания*[1]:

$$D(X) = M[X - M(X)]^2 \qquad \textbf{(3.10)}$$

или $$D(X) = M(X - a)^2, \text{ где } a = M(X).$$

В качестве характеристики рассеяния нельзя брать математическое ожидание отклонения случайной величины от ее математического ожидания $M(X-a)$, ибо согласно свойству **6** математического ожидания эта величина равна нулю для любой случайной величины.

Выбор дисперсии, определяемой по формуле (**3.10**), в качестве характеристики рассеяния значений случайной величины X оправдывается также тем, что, как можно показать, *математическое ожидание квадрата отклонения случайной величины* X *от постоянной величины* C *минимально именно тогда, когда эта постоянная* C *равна математическому ожиданию* $M(X) = a$, т.е.

$$\min_C M(X-C)^2 = M(X-a)^2 = D(X).$$

Если случайная величина X — дискретная с конечным числом значений, то

$$D(X) = \sum_{i=1}^{n} (x_i - a)^2 p_i. \qquad \textbf{(3.11)}$$

Если случайная величина X — дискретная с бесконечным, но счетным множеством значений, то

---

[1] Для дисперсии случайной величины X в литературе используется также обозначение $\mathrm{var}(X)$.

$$D(X) = \sum_{i=1}^{\infty} (x_i - a)^2 p_i. \qquad (3.12)$$

(если ряд в правой части равенства сходится).

Дисперсия $D(X)$ имеет размерность квадрата случайной величины, что не всегда удобно. Поэтому в качестве показателя рассеяния используют также величину $\sqrt{D(X)}$.

О п р е д е л е н и е. ***Средним квадратическим отклонением*** (***стандартным отклонением*** или ***стандартом***) σ_x *случайной величины X называется арифметическое значение корня квадратного из ее дисперсии*:

$$\sigma_x = \sqrt{D(X)}. \qquad (3.13)$$

▷ **Пример 3.8.** В задаче о стрелках (см. § 3.3) вычислить дисперсию и среднее квадратическое отклонение числа выбитых очков для каждого стрелка.

Р е ш е н и е. В примере **3.5** были вычислены $M(X)=5{,}36$ и $M(Y)=5{,}36$. По формулам(**3.12**) и (**3.13**)

$$D(X) = (0 - 5{,}36)^2 \cdot 0{,}15 + (1 - 5{,}36)^2 \cdot 0{,}11 + \ldots +$$

$$+ (10 - 5{,}36)^2 \cdot 0{,}20 = 13{,}61,$$

$$\sigma_x = \sqrt{D(X)} = 3{,}69;$$

$$D(Y) = (0 - 5{,}36)^2 \cdot 0{,}01 + (1 - 5{,}36)^2 \cdot 0{,}03 + \ldots +$$

$$+ (10 - 5{,}36)^2 \cdot 0{,}02 = 4{,}17,$$

$$\sigma_y = \sqrt{D(Y)} = 2{,}04.$$

Итак, при равенстве средних значений числа выбиваемых очков ($M(X)=M(Y)$) его дисперсия, т.е. характеристика рассеяния относительно среднего значения, меньше у второго стрелка ($D(X)<D(Y)$) и, очевидно, ему для получения более высоких результатов стрельбы по сравнению с первым стрелком нужно сместить «центр» распределения числа выбиваемых очков, т.е. увеличить $M(Y)$, научившись лучше целиться в мишень. ◀

Отметим с в о й с т в а дисперсии случайной величины.

1. *Дисперсия постоянной величины равна нулю*:

$$D(C) = 0. \qquad \textbf{(3.14)}$$

□ $D(C) = M[C - M(C)]^2 = M(C - C)^2 = M(0) = 0.$ ■

2. *Постоянный множитель можно выносить за знак дисперсии, возведя его при этом в квадрат*:

$$D(kX) = k^2 D(X). \qquad \textbf{(3.15)}$$

□ Учитывая свойство **2** математического ожидания, получим

$$D(kX) = M[kX - M(kX)]^2 = M[kX - kM(X)]^2 =$$

$$= k^2 M[X - M(X)]^2 = k^2 D(X).$$ ■

3. *Дисперсия случайной величины равна разности между математическим ожиданием квадрата случайной величины и квадратом ее математического ожидания*:

$$D(X) = M(X^2) - [M(X)]^2, \qquad \textbf{(3.16)}$$

или

$$D(X) = M(X^2) - a^2, \text{ где } a = M(X).$$

□ Пусть $M(X)=a$. Тогда $D(X)=M(X-a)^2=M(X^2-2aX+a^2)$.

Учитывая, что a — величина постоянная, неслучайная[1], найдем

$$D(X)=M(X)^2-2aM(X)+a^2=M(X^2)-2a\cdot a+a^2 = M(X^2)-a^2.$$ ■

Это свойство часто используют при вычислении дисперсии. Вычисление по формуле **(3.16)** дает, например, упрощение расчетов по сравнению с основной формулой **(3.11)**, если значения x_i случайной величины — целые, а математическое ожидание, а значит, и разности $(x_i - a)$ — нецелые числа.

▷ **Пример 3.9.** По данным примера **3.7** (задачи о стрелках) вычислить дисперсии случайных величин X, Y, используя свойство **3**.

---

[1] Из определения как математического ожидания $M(X)$, так и дисперсии $D(X)$ случайной величины X следует, что сами $M(X)$ и $D(X)$ — величины *неслучайные, постоянные.*

Р е ш е н и е. Вначале найдем

$$M\left(X^2\right)=\sum_{i=1}^{n} x_i^2 p_i = 0^2\cdot 0{,}15+1^2\cdot 0{,}10+\ldots+9^2\cdot 0{,}12+10^2\cdot 0{,}20=42{,}34.$$

Теперь по формуле **(3.16)**

$$D(X) = M(X^2) - a^2 = 42{,}34 - 5{,}36^2 = 13{,}61.$$

Аналогично $D(Y) = 4{,}17$. ▶

4. *Дисперсия алгебраической суммы конечного числа независимых случайных величин равна сумме их дисперсий*[1]:

$$D(X\pm Y) = D(X)+D(Y). \qquad \textbf{(3.17)}$$

□ По свойству **3**:

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

Обозначая $M(X)=a_x$, $M(Y)=a_y$ и учитывая, что для независимых случайных величин $M(XY)=M(X)M(Y)$, получим

$$D(X\pm Y)=M(X^2) \pm 2a_xa_y + M(Y^2)- a_x^2 \mp 2a_xa_y- a_y^2 =$$

$$=[M(X^2) - a_x^2]+[M(Y^2)- a_y^2]=D(X)+D(Y).$$ ■

Обращаем внимание на то, что *дисперсия как суммы, так и разности независимых случайных величин X и Y равна с у м м е их дисперсий*, т.е.

$$D(X+Y)=D(X-Y)=D(X)+D(Y).$$

▷ **Пример 3.10.** Найти дисперсию случайной величины $Z=8X-5Y+7$, если известно, что случайные величины X и Y независимы и $D(X) = 1{,}5$, $D(Y) = 1$.

Р е ш е н и е. Используя свойства **1**, **2**, **4** дисперсии, найдем

$$D(Z)=8^2 D(X)+5^2 D(Y)+0=64\cdot 1{,}5+25\cdot 1=121.$$ ▶

---

[1] Записываем свойство для двух случайных величин; случай зависимых случайных величин рассматривается в § 5.6 — см. **(5.42)**, **(5.43)**.

Если использовать механическую интерпретацию распределения случайной величины, то ее дисперсия представляет собой момент инерции распределения масс относительно центра масс (математического ожидания).

Замечание. Обратим внимание на интерпретацию математического ожидания и дисперсии в финансовом анализе. Пусть, например, известно распределение доходности X некоторого актива (например, акции), т.е. известны значения доходности x_i и соответствующие их вероятности p_i за рассматриваемый промежуток времени. Тогда, очевидно, математическое ожидание $M(X)$ выражает *среднюю (прогнозную) доходность актива*, а дисперсия $D(X)$ или среднее квадратическое отклонение σ_x — меру отклонения, колеблемости доходности от ожидаемого среднего значения, т.е. *риск* данного актива.

Математическое ожидание, дисперсия, среднее квадратическое отклонение и другие *числа, призванные в сжатой форме выразить наиболее существенные черты распределения, называются* ***числовыми характеристиками*** *случайной величины.*

Обращаем внимание на то, что сама величина X — *случайная*, а ее *числовые характеристики являются величинами неслучайными, постоянными.*

В теории вероятностей числовые характеристики играют большую роль. Часто удается решать вероятностные задачи, оперируя лишь числовыми характеристиками случайных величин. Применение вероятностных методов для решения практических задач в значительной мере определяется умением пользоваться числовыми характеристиками случайной величины, оставляя в стороне законы распределения.

3.5. Функция распределения случайной величины

До сих пор в качестве исчерпывающего описания дискретной случайной величины мы рассматривали закон ее распределения, представляющий собой ряд распределения или формулу, позволяющие находить вероятности любых значений случайной величины X. Однако такое описание случайной величины X не является единственным, а, главное, не универсально. Так, оно неприменимо для непрерывной случайной величины, так как, во-первых, нельзя перечислить все бесконечное несчетное множество ее значений; во-вторых, как мы увидим дальше, вероят-

ности каждого отдельно взятого значения непрерывной случайной величины равны нулю.

Для описания закона распределения случайной величины X возможен и другой подход: рассматривать не вероятности событий $X=x$ для разных x (как это имеет место в ряде распределения), а вероятности события $X<x$, где x — текущая переменная. Вероятность $P(X<x)$, очевидно, зависит от x, т.е. является некоторой функцией от x.

О п р е д е л е н и е. ***Функцией распределения*** *случайной величины X называется функция $F(x)$, выражающая для каждого x вероятность того, что случайная величина X примет значение, меньшее x:*

$$F(x) = P(X < x). \qquad \textbf{(3.18)}$$

Функцию $F(x)$ иногда называют *интегральной функцией распределения* или *интегральным законом распределения.*

Геометрически функция распределения интерпретируется как вероятность того, что случайная точка X попадет левее заданной точки x (рис. 3.4)

Рис. 3.4

▷ **Пример 3.11.** Дан ряд распределения случайной величины

X:

x_i	1	4	5	7
p_i	0,4	0,1	0,3	0,2

Найти и изобразить графически ее функцию распределения.

Р е ш е н и е. Будем задавать различные значения x и находить для них $F(x) = P(X < x)$.

1. Если $x \le 1$, то, очевидно, $F(x) = 0$ (в том числе и при $x = 1$ $F(1) = P(x < 1) = 0$).

2. Пусть $1 < x \le 4$ (например, $x = 2$);
$F(x) = P(X = 1) = 0{,}4$. Очевидно, что и $F(4) = P(X < 4) = 0{,}4$.

3. Пусть $4 < x \le 5$ (например, $x = 4{,}25$);

$$F(x)=P(X<x)=P(X=1)+P(X=4)=0{,}4+0{,}1=0{,}5.$$

Очевидно, что и $F(5) = 0{,}5$.

4. Пусть $5 < x \le 7$. $F(x)=\left[P(X=1)+P(X=4)\right]+P(X=5)=$ $=0{,}5+0{,}3=0{,}8$. Очевидно, что и $F(7) = 0{,}8$,

5. Пусть $x > 7$. $F(x)=\left[P(X=1)+P(X=4)+P(X=5)\right]+P(X=7)=$ $=0{,}8+0{,}2=1$.

Изобразим функцию $F(x)$ графически (рис. 3.5).

Итак,
$$F(x)=\begin{cases}0 & \text{при } x\le 1,\\ 0{,}4 & \text{при } 1<x\le 4,\\ 0{,}5 & \text{при } 4<x\le 5,\\ 0{,}8 & \text{при } 5<x\le 7,\\ 1{,}0 & \text{при } x>7.\end{cases}$$

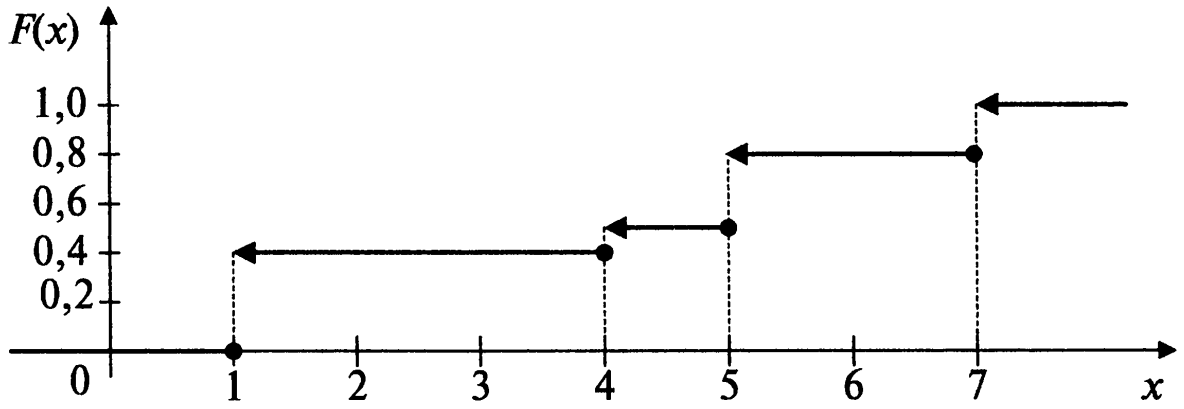


Рис. 3.5

Заметим, что при подходе слева к точкам разрыва функция сохраняет свое значение (про такую функцию говорят, что она непрерывна слева). Эти точки на графике выделены. ▶

Этот пример позволяет прийти к утверждению, что *функция распределения любой дискретной случайной величины есть разрывная ступенчатая функция, скачки которой происходят в точках, соответствующих возможным значениям случайной величины и равны вероятностям этих значений.* Сумма всех скачков функции $F(x)$ равна 1.

Рассмотрим общие с в о й с т в а функции распределения.

1. *Функция распределения случайной величины есть неотрицательная функция, заключенная между нулем и единицей:*

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

□ Утверждение следует из того, что функция распределения — это вероятность. ■

2. *Функция распределения случайной величины есть неубывающая функция на всей числовой оси.*

□ Пусть x_1 и x_2 — точки числовой оси, причем $x_2 > x_1$. Покажем, что $F(x_2) \ge F(x_1)$. Рассмотрим два несовместных события $A = (X < x_1)$, $B = (x_1 \le X < x_2)$. Тогда $A + B = (X < x_2)$.

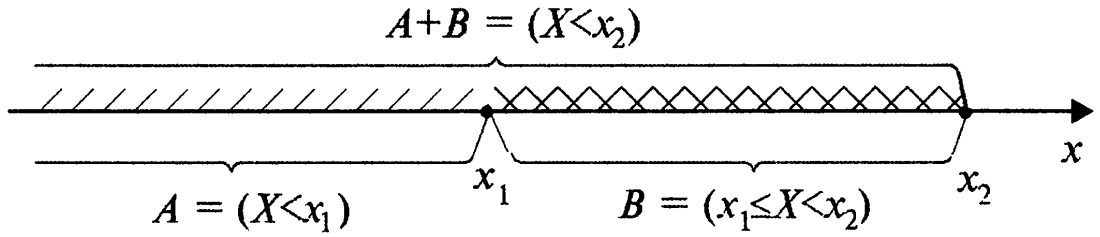


Рис. 3.6

Это соотношение между событиями легко усматривается из их геометрической интерпретации (рис. 3.6). По теореме сложения

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

или $$P(X < x_2) = P(X < x_1) + P(x_1 \le X < x_2),$$

откуда

$$F(x_2) = F(x_1) + P(x_1 \le X < x_2). \qquad \textbf{(3.19)}$$

Так как вероятность $P(x_1 \le X < x_2) \ge 0$, то $F(x_2) \ge F(x_1)$, т.е. $F(x)$ — неубывающая функция. ■

3. *На минус бесконечности функция распределения равна нулю, на плюс бесконечности равна единице*, т.е.

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0, \qquad F(+\infty) = \lim_{x \to +\infty} F(x) = 1.$$

□ $F(-\infty) = P(X < -\infty) = 0$ как вероятность невозможного события $X < -\infty$.

$F(+\infty) = P(X < +\infty) = 1$ как вероятность достоверного события $X < +\infty$. ■

4. *Вероятность попадания случайной величины в интервал* $[x_1, x_2)$ (включая x_1) *равна приращению ее функции распределения на этом интервале*, т.е.

$$P(x_1 \le X < x_2) = F(x_2) - F(x_1). \qquad \textbf{(3.20)}$$

□ Формула **(3.20)** следует непосредственно из формулы **(3.19)**. ■

▷ **Пример 3.12.** Функция распределения случайной величины X имеет вид:

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

Найти вероятность того, что случайная величина примет значение в интервале [1;3).

Р е ш е н и е. По формуле **(3.20)**

$$P(1 \le X < 3) = F(3) - F(1) = 1 - \frac{1}{2} = \frac{1}{2}. \blacktriangleright$$

3.6. Непрерывные случайные величины. Плотность вероятности

Выше дано понятие непрерывной случайной величины, имеющей бесконечное несчетное множество значений. Приведем теперь более строгое определение.

О п р е д е л е н и е. *Случайная величина X называется* ***непрерывной****, если ее функция распределения непрерывна в любой точке и дифференцируема всюду, кроме, быть может, отдельных точек.*

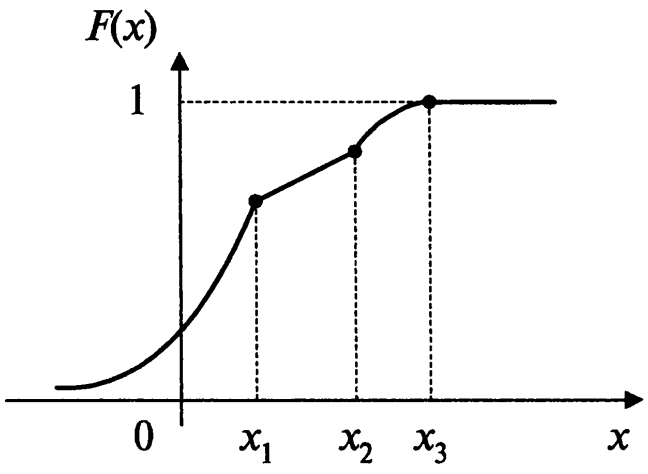


Рис. 3.7

На рис. 3.7 показана функция распределения непрерывной случайной величины X, дифференцируемая во всех точках, кроме трех точек излома.

Теорема. *Вероятность любого отдельно взятого значения непрерывной случайной величины равна нулю*[1].

□ Покажем, что для любого значения x_1 случайной величины X вероятность $P(X = x_1) = 0$. Представим $P(X = x_1)$ в виде $P(X = x_1) = \lim\limits_{x_2 \to x_1} P(x_1 \le X < x_2)$.

Применяя свойство **(3.20)** функции распределения случайной величины X и учитывая непрерывность $F(x)$, получим

$$P(X = x_1) = \lim_{x_2 \to x_1} \left[F(x_2) - F(x_1)\right] =$$

$$= \lim_{x_2 \to x_1} F(x_2) - F(x_1) = F(x_1) - F(x_1) = 0. \blacksquare$$

До сих пор мы рассматривали испытания, сводившиеся к схеме случаев, и нулевой вероятностью обладали лишь невозможные события. Из приведенной выше теоремы следует, что нулевой вероятностью могут обладать и возможные события, так как событие, состоящее в том, что случайная величина X приняла конкретное значение x_1, является возможным. На первый взгляд этот вывод может показаться парадоксальным. Действительно, если, например, событие $\alpha \le X \le \beta$ имеет отличную от нуля вероятность, то оказывается, что оно представляет собой сумму событий, состоящих в принятии случайной величиной X любых конкретных значений на отрезке $[\alpha, \beta]$ и имеющих нулевую вероятность. Но никакого противоречия здесь нет, ибо теорема сложения (точнее, аксиома сложения — см. § 1.12) справедлива только для конечного и счетного бесконечного множества событий, а множество событий, обозначающих отмеченную сумму, таковым не является.

Представление о событии, имеющем отличную от нуля вероятность, но складывающемся из событий с нулевой вероятностью, не более парадоксально, чем представление об отрезке, имеющем определенную длину, тогда как ни одна точка отрезка отличной от нуля длиной не обладает. Отрезок состоит из таких точек, но его длина не равна сумме их длин.

Далее, рассматривая теорему Бернулли (см. § 6.4), мы убедимся в том, что при $n \to \infty$ частость события m/n приближает-

---

[1] Поэтому непрерывную случайную величину можно было определить и иначе: случайная величина *непрерывна*, если вероятность любого отдельно взятого ее значения равна нулю.

ся к вероятности этого события. Поэтому из того, что вероятность события равна нулю, следует только, что при неограниченном повторении опыта его частота m/n будет приближаться к нулю, т.е. событие будет появляться сколь угодно редко.

Следствие. *Если X — непрерывная случайная величина, то вероятность попадания случайной величины в интервал (x_1,x_2) не зависит от того, является этот интервал открытым или закрытым,* т.е.

$$P(x_1 < X < x_2) = P(x_1 \le X < x_2) = P(x_1 < X \le x_2) = P(x_1 \le X \le x_2).$$

□ $$P(x_1 \le X \le x_2) = P(X = x_1) + P(x_1 < X < x_2) + P(X = x_2) =$$
$$= 0 + P(x_1 < X < x_2) + 0 = P(x_1 < X < x_2).$$

Аналогично доказываются и другие равенства. ■

Задание непрерывной случайной величины с помощью функции распределения не является единственным. Введем понятие плотности вероятности непрерывной случайной величины.

Рассмотрим вероятность попадания непрерывной случайной величины на участок $[x, x+\Delta x]$. По формуле **(3.20)** вероятность

$$P(x \le X \le x + \Delta x) = F(x + \Delta x) - F(x),$$

т.е. равна приращению функции распределения $F(x)$ на этом участке. Тогда вероятность, приходящаяся на единицу длины, т.е. средняя плотность вероятности на участке от x до $x + \Delta x$ равна

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

Переходя к пределу при $\Delta x \to 0$, получим плотность вероятности в точке x:

$$\lim_{\Delta x \to 0} \frac{P(x \le X \le x + \Delta x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = F'(x),$$

представляющую производную функции распределения $F(x)$ (напомним, что для непрерывной случайной величины $F(x)$ — дифференцируемая функция).

Определение. ***Плотностью вероятности*** (***плотностью распределения*** или просто ***плотностью***) $\varphi(x)$ *непрерывной случайной величины X называется производная ее функции распределения*

$$\varphi(x) = F'(x). \qquad (3.21)$$

Про случайную величину X говорят, что она имеет распределение (распределена) с плотностью $\varphi(x)$ на определенном участке оси абсцисс.

Плотность вероятности $\varphi(x)$, как и функция распределения $F(x)$, является одной из форм закона распределения, но в отличие от функции распределения она существует только для непрерывных случайных величин.

Плотность вероятности иногда называют *дифференциальной функцией* или *дифференциальным законом распределения.*

График плотности вероятности $\varphi(x)$ называется *кривой распределения.*

▷ **Пример 3.13.** По данным примера **3.12** найти плотность вероятности случайной величины X.

Решение. Плотность вероятности $\varphi(x) = F'(x)$, т.е.

$$\varphi(x) = \begin{cases} 0 & \text{при } x \le 0 \text{ и } x > 2, \\ 1/2 & \text{при } 0 < x \le 2. \end{cases}$$ ▶

Отметим свойства плотности вероятности непрерывной случайной величины.

1. *Плотность вероятности—неотрицательная функция*, т.е.

$$\varphi(x) \ge 0.$$

□ $\varphi(x) \ge 0$ как производная монотонно неубывающей функции $F(x)$. ■

2. *Вероятность попадания непрерывной случайной величины в интервал $[a, b]$ равна определенному интегралу от ее плотности вероятности в пределах от a до b*, т.е.

$$P(a \le X \le b) = \int_a^b \varphi(x)dx. \qquad (3.22)$$

□ Согласно свойству **4** функции распределения

$$P(a \le X \le b) = F(b) - F(a).$$

Так как $F(x)$ есть первообразная для плотности вероятности $\varphi(x)$ (ибо $F'(x) = \varphi(x)$), то по формуле Ньютона—Лейбница приращение первообразной на отрезке $[a, b]$ есть определенный интеграл $\int_a^b \varphi(x)dx$, т.е. формула (**3.20**) верна. ■

Геометрически полученная вероятность равна площади фигуры, ограниченной сверху кривой распределения и опирающейся на отрезок $[a,b]$ (рис. 3.8).

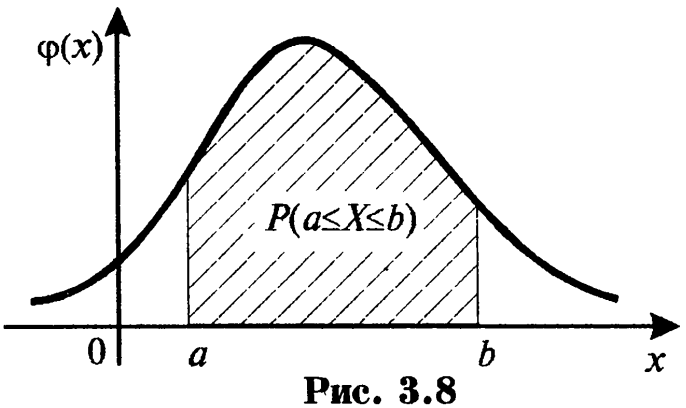


Рис. 3.8

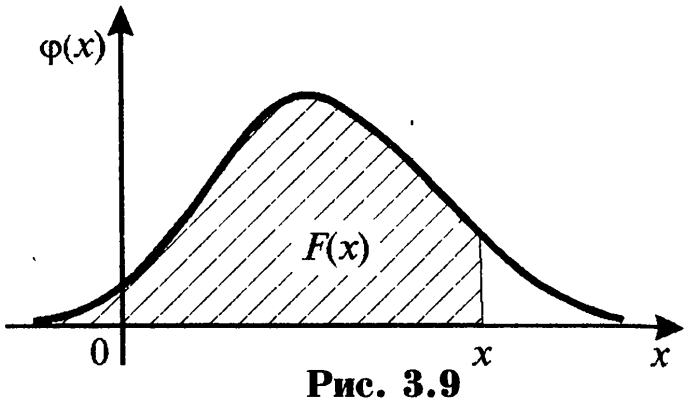


Рис. 3.9

3. *Функция распределения непрерывной случайной величины может быть выражена через плотность вероятности по формуле:*

$$F(x)=\int_{-\infty}^{x}\varphi(x)dx. \qquad \textbf{(3.23)}$$

Формула **(3.23)** получается из формулы **(3.22)** при $a\to-\infty$, если верхний предел b заменить на переменный предел x.

Геометрически функция распределения равна площади фигуры, ограниченной сверху кривой распределения и лежащей левее точки x (рис. 3.9).

4. *Несобственный интеграл в бесконечных пределах от плотности вероятности непрерывной случайной величины равен единице:*

$$\int_{-\infty}^{+\infty}\varphi(x)dx=1. \qquad \textbf{(3.24)}$$

□ По формуле **(3.23)**: $F(x)=\int_{-\infty}^{x}\varphi(x)dx$ и при $x\to+\infty$ $F(+\infty)=1$,

т.е. верно равенство **(3.24)**. ■

Геометрически свойства **1** и **4** плотности вероятности означают, что ее график — *кривая распределения — лежит не ниже оси абсцисс, и полная площадь фигуры, ограниченной кривой распределения и осью абсцисс, равна единице.*

Понятие математического ожидания $M(X)$ и дисперсии $D(X)$, введенные выше (§ 3.3, 3.4) для дискретной случайной величины, можно распространить на непрерывные случайные величины.

Для получения соответствующих формул для $M(X)$ и $D(X)$ достаточно в формулах **(3.3)** и **(3.11)** для дискретной случайной

величины X заменить знак суммирования $\sum_{i=1}^{n}$ по всем ее значениям знаком интеграла с бесконечными пределами $\int_{-\infty}^{+\infty}$, «скачущий» аргумент x_i — непрерывно меняющимся x, а вероятность p_i — элементом вероятности $\varphi(x)dx$. Под *элементом вероятности* понимается вероятность попадания случайной величины X на участок $[x, x + dx]$ (с точностью до бесконечно малых более высоких порядков); геометрически элемент вероятности приближенно равен площади элементарного прямоугольника, опирающегося на отрезок $[x, x + dx]$ (рис. 3.10).

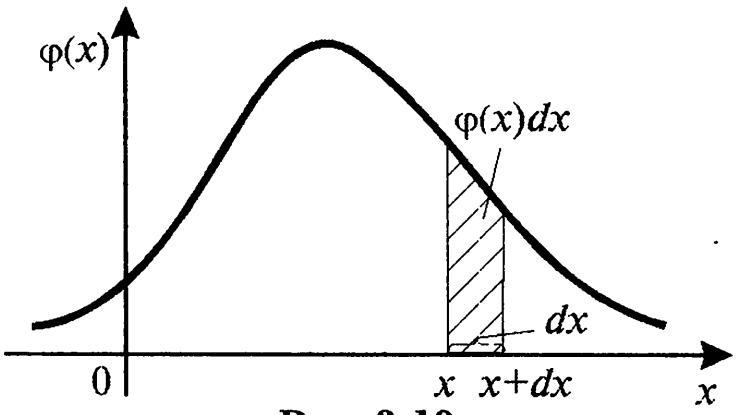


Рис. 3.10

В результате получим следующие формулы для математического ожидания и дисперсии непрерывной случайной величины X:

$$a = M(X) = \int_{-\infty}^{+\infty} x\,\varphi(x)dx \qquad \textbf{(3.25)}$$

(если интеграл абсолютно сходится) и

$$D(X) = \int_{-\infty}^{+\infty} (x - a)^2\,\varphi(x)dx \qquad \textbf{(3.26)}$$

(если интеграл сходится).

На практике обычно область значений случайной величины, для которых $\varphi(x) \neq 0$, ограничена и указанные интегралы сходятся, а значит, существуют $M(X)$ и $D(X)$.

Все свойства математического ожидания и дисперсии, рассмотренные выше для дискретных случайных величин, справедливы и для непрерывных величин[1].

---

[1] Заметим, что сохраняет тот же смысл механическая интерпретация математического ожидания как абсциссы центра масс для единичной массы, распределенной в данном случае непрерывно на оси абсцисс с плотностью вероятности $\varphi(x)$, и дисперсии как момента инерции распределения масс относительно центра масс.

В частности, свойство **3** дисперсии (формула (**3.16**)) имеет вид:

$$D(X)=M\left(X^2\right)-a^2 \text{ или } D(X)=\int_{-\infty}^{+\infty}x^2\varphi(x)dx-a^2. \qquad (\mathbf{3.27})$$

З а м е ч а н и е. Наряду с дискретными и непрерывными случайными величинами на практике встречаются *смешанные* случайные величины, для которых функция распределения $F(x)$ на некоторых участках непрерывна, а в отдельных точках имеет разрывы. Примером смешанной случайной величины может служить заработок рабочего, пропорциональный его выработке, но не меньший гарантированного размера оплаты x_0. (При $x = x_0$ функция распределения $F(x)$ имеет скачок от нуля до некоторого значения p_0, а при $x > x_0$ непрерывно возрастает.) Для смешанных случайных величин остается справедливой формула (**3.20**) вероятности попадания случайной величины на любой интервал $[x_0, x_1)$.

▷ **Пример 3.14.** Функция $\varphi(x)$ задана в виде:

$$\varphi(x)=\begin{cases}0 & \text{при } x\le 1,\\ \dfrac{A}{x^4} & \text{при } x>1.\end{cases}$$

Найти: а) значение постоянной A, при которой функция будет плотностью вероятности некоторой случайной величины X; б) выражение функции распределения $F(x)$; в) вычислить вероятность того, что случайная величина X примет значение на отрезке $[2;3]$; г) найти математическое ожидание и дисперсию случайной величины X.

Р е ш е н и е. а) Для того чтобы $\varphi(x)$ была плотностью вероятности некоторой случайной величины X, она должна быть неотрицательна, т.е. $\varphi(x)\ge 0$ или $\dfrac{A}{x^4}\ge 0$, откуда $A\ge 0$, и она должна удовлетворять свойству **4**. Поэтому в соответствии с формулой (**3.24**) $\int_{-\infty}^{+\infty}\varphi(x)\,dx=1$.

Следовательно,

$$\int_{-\infty}^{+\infty}\varphi(x)\,dx=\int_{-\infty}^{1}0\cdot dx+\int_{1}^{+\infty}\frac{A}{x^4}dx=0+\lim_{b\to+\infty}\int_{1}^{b}\frac{A}{x^4}dx=$$

$$=\frac{A}{3}\lim_{b\to+\infty}\left(-\frac{1}{x^3}\Bigg|_1^b\right)=\frac{A}{3}\lim_{b\to+\infty}\left(1-\frac{1}{b^3}\right)=\frac{A}{3}=1,$$

откуда $A = 3$.

б) По формуле (**3.23**) найдем $F(x)$.

Если $x \le 1$, то $F(x) = \int_{-\infty}^{x} \varphi(x)\,dx = \int_{-\infty}^{x} 0 \cdot dx = 0.$

Если $x > 1$, то $F(x) = 0 + \int_{1}^{x} \frac{3}{x^4}\,dx = -\frac{1}{x^3}\Big|_1^x = 1 - \frac{1}{x^3}.$

Таким образом, $F(x) = \begin{cases} 0 & \text{при } x \le 1, \\ 1 - \frac{1}{x^3} & \text{при } x > 1. \end{cases}$

в) По формуле (**3.22**)

$$P(2 \le X \le 3) = \int_{2}^{3} \frac{3}{x^4}\,dx = -\frac{1}{x^3}\Big|_2^3 = \frac{1}{2^3} - \frac{1}{3^3} = \frac{19}{216}.$$

Вероятность $P(2 \le X \le 3)$ можно было найти непосредственно как приращение функции распределения по формуле (**3.19**):

$$P(2 \le X \le 3) = F(3) - F(2) = \left(1 - \frac{1}{3^3}\right) - \left(1 - \frac{1}{2^3}\right) = \frac{19}{216}.$$

г) По формуле (**3.25**) вычислим

$$a = M(X) = \int_{-\infty}^{+\infty} x\varphi(x)dx = \int_{-\infty}^{1} 0 \cdot dx + \int_{1}^{+\infty} x\left(\frac{3}{x^4}\right)dx = 0 + 3\lim_{b\to+\infty}\int_1^b \frac{dx}{x^3} =$$

$$= 3\lim_{b\to+\infty}\left(-\frac{1}{2x^2}\Big|_1^b\right) = \frac{3}{2}\lim_{b\to+\infty}\left(1 - \frac{1}{b^2}\right) = \frac{3}{2}.$$

Дисперсию $D(X)$ вычислим по формуле (**3.27**). Вначале найдем

$$M(X^2) = \int_{-\infty}^{+\infty} x^2\varphi(x)dx = \int_{-\infty}^{+\infty} x^2\left(\frac{3}{x^4}\right)dx = 3$$

(вычисление интеграла аналогично приведенному выше). Теперь

$$D(X) = 3 - \left(\frac{3}{2}\right)^2 = \frac{3}{4}. \blacktriangleright$$

Замечание. В ряде случаев, если имеется график функции распределения $F(x)$, полезно иметь в виду геометрическую интерпретацию математического ожидания $M(X)$ случайной величины X:

$$M(X) = S_2 - S_1,$$

где S_2 и S_1 — площади фигур, заключенных соответственно между осью Oy, прямой $y = 1$ и кривой $y = F(x)$ на интервале

$(0;+\infty)$ и между кривой $y = F(x)$ и осями Ox и Oy на промежутке $(-\infty;0)$ (рис. 3.11).

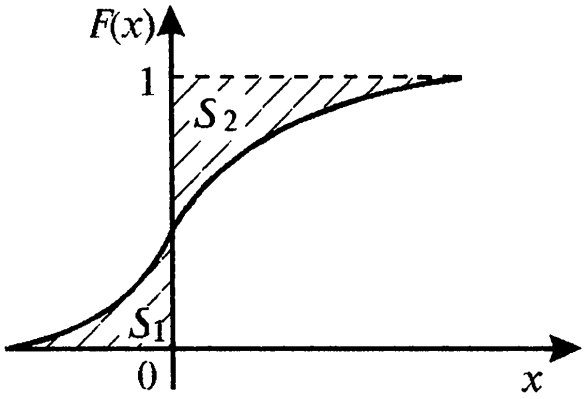


Рис. 3.11

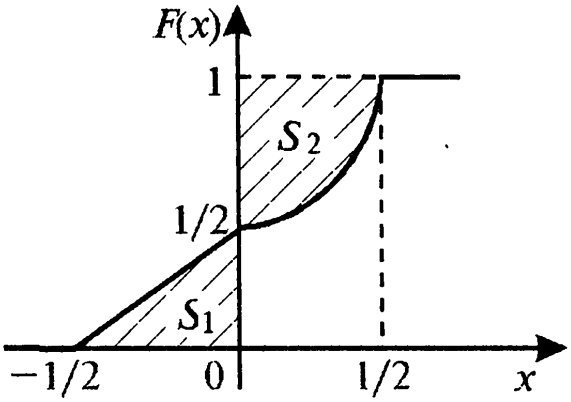


Рис. 3.12

Так, например, для нахождения математического ожидания $M(X)$ случайной величины X, заданной функцией распределения $F(x)$, состоящей из участков прямых и дуги окружности (рис. 3.12), нет необходимости находить $\varphi(x)$ по формуле (**3.21**), а затем $M(X)$ по формуле (**3.25**). Значительно проще найти $M(X)$, используя его геометрическую интерпретацию, т.е.

$$M(X) = S_2 - S_1 = \frac{1}{4}\pi R^2 - \frac{1}{2}ah = \frac{1}{4}\pi\left(\frac{1}{2}\right)^2 - \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{\pi-2}{16} \approx 0{,}072.$$

3.7. Мода и медиана. Квантили. Моменты случайных величин. Асимметрия и эксцесс

Кроме математического ожидания и дисперсии, в теории вероятностей применяется еще ряд числовых характеристик, отражающих те или иные особенности распределения.

О п р е д е л е н и е. ***Модой*** $Mo(X)$ *случайной величины* X *называется ее наиболее вероятное значение* (для которого вероятность p_i или плотность вероятности $\varphi(x)$ достигает максимума).

Если вероятность или плотность вероятности достигает максимума не в одной, а в нескольких точках, распределение называется *полимодальным* (рис. 3.13).

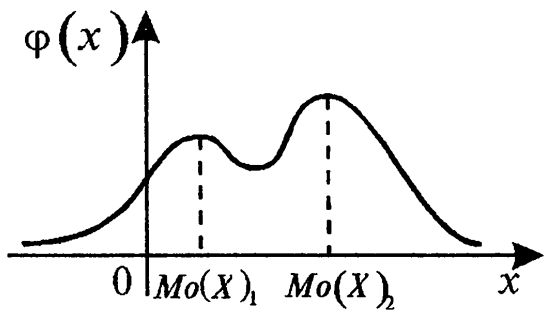


Рис. 3.13

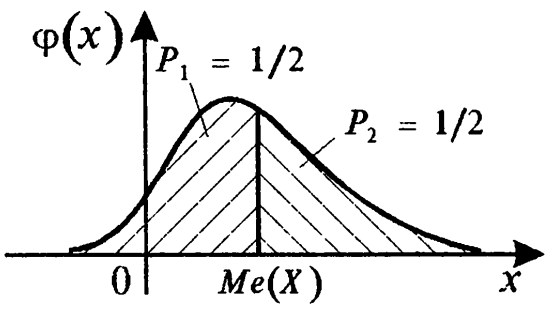


Рис. 3.14

Определение. ***Медианой*** *Me(X) непрерывной случайной величины X называется такое ее значение, для которого*

$$P\big(X < Me(X)\big) = P\big(X > Me(X)\big) = \frac{1}{2}, \qquad \textbf{(3.28)}$$

т.е. вероятность того, что случайная величина X примет значение, меньшее медианы $Me(X)$ или большее ее, одна и та же и равна 1/2. Геометрически вертикальная прямая $x=Me(X)$, проходящая через точку с абсциссой, равной $Me(X)$, делит площадь фигуры под кривой распределения на две равные части (рис. 3.14). Очевидно, что в точке $x=Me(X)$ функция распределения равна 1/2, т.е. $F\big(Me(X)\big) = 1/2$ (рис. 3.15).

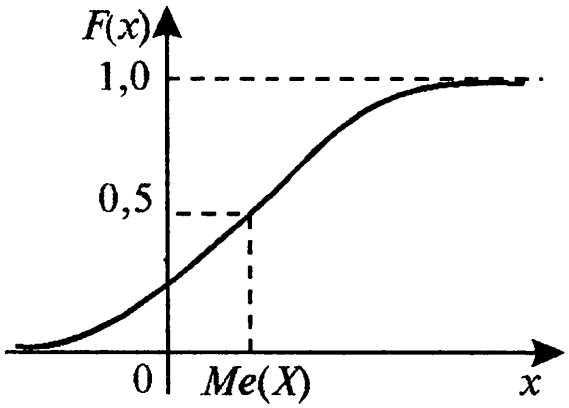


Рис. 3.15

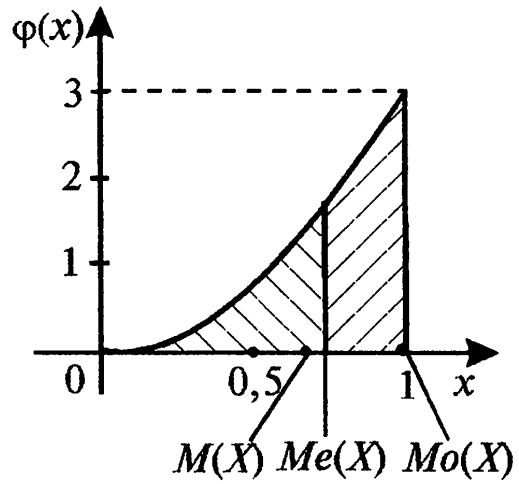


Рис. 3.16

▷ **Пример 3.15.** Найти моду, медиану и математическое ожидание случайной величины X с плотностью вероятности $\varphi(x) = 3x^2$ при $x \in [0;1]$.

Р е ш е н и е. Кривая распределения представлена на рис. 3.16. Очевидно, что плотность вероятности $\varphi(x)$ максимальна при $x=Mo(X)=1$.

Медиану $Me(X)=b$ найдем из условия **(3.28)**:

$$\int_{-\infty}^{b} \varphi(x)dx = \frac{1}{2}$$

или

$$\int_{-\infty}^{b} \varphi(x)dx = \int_{-\infty}^{0} 0 \cdot dx + \int_{0}^{b} 3x^2 dx = x^3 \Big|_0^b = b^3 = \frac{1}{2},$$

откуда

$$b = Me(X) = \sqrt[3]{1/2} \approx 0{,}79.$$

Математическое ожидание вычислим по формуле (**3.25**):

$$M(X)=\int_{-\infty}^{+\infty}x\varphi(x)dx=\int_{-\infty}^{0}0\cdot dx+\int_{0}^{1}x\left(3x^2\right)dx+\int_{1}^{+\infty}0\cdot dx=\frac{3}{4}x^4\Big|_0^1=0{,}75.$$

Взаимное расположение точек $M(X)$, $Me(X)$ и $Mo(X)$ в порядке возрастания абсцисс показано на рис. 3.16. ▷

Наряду с отмеченными выше числовыми характеристиками для описания случайной величины используется понятие квантилей и процентных точек.

О п р е д е л е н и е. ***Квантилем уровня*** q (или ***q-квантилем***) *называется такое значение* x_q *случайной величины, при котором функция ее распределения принимает значение, равное* q, т.е.

$$F(x_q)=P(X<x_q)=q. \qquad (3.29)$$

Некоторые квантили получили особое название. Очевидно, что введенная выше *медиана* случайной величины есть квантиль уровня 0,5, т.е. $Me(X)=x_{0,5}$. Квантили $x_{0,25}$ и $x_{0,75}$ получили название соответственно *верхнего* и *нижнего квантилей*[1].

С понятием квантиля тесно связано понятие ***процентной точки***. Под $100q\%$*-ной точкой* подразумевается квантиль x_{1-q}, т.е. такое значение случайной величины X, при котором $P\left(X\geq x_{1-q}\right)=q$.

▷**Пример 3.16.** По данным примера **3.15** найти квантиль $x_{0,3}$ и 30%-ную точку случайной величины X.

Р е ш е н и е. По формуле (**3.23**) функция распределения

$$F(x)=\int_{-\infty}^{x}\varphi(x)dx=\int_{-\infty}^{0}0\cdot dx+\int_{0}^{x}3x^2dx=x^3.$$

Квантиль $x_{0,3}$ найдем из уравнения (**3.29**), т.е. $x_{0,3}^3=0{,}3$, откуда $x_{0,3}\approx 0{,}67$. Найдем 30%-ную точку случайной величины X, или квантиль $x_{0,7}$, из уравнения $x_{0,7}^3=0{,}7$, откуда $x_{0,7}\approx 0{,}89$. ▷

Среди числовых характеристик случайной величины особое значение имеют м о м е н т ы — начальные и центральные.

О п р е д е л е н и е. ***Начальным моментом*** *k-го порядка случайной величины X называется математическое ожидание k-й степени этой величины*:

---

[1] В литературе встречаются также термины: *децили* (под которыми понимаются квантили $x_{0,1}$, $x_{0,2}$,..., $x_{0,9}$) и *процентили* (квантили $x_{0,01}$, $x_{0,02}$,..., $x_{0,99}$).

$$\nu_k = M(X^k). \qquad (3.30)$$

Определение. ***Центральным моментом*** *k-го порядка случайной величины X называется математическое ожидание k-й степени отклонения случайной величины X от ее математического ожидания:*

$$\mu_k = M\left[X - M(X)\right]^k, \qquad (3.31)$$

или $\mu_k = M(X-a)^k$, где $a = M(X)$.

Формулы для вычисления моментов для дискретных случайных величин (принимающих значения x_i с вероятностями p_i) и непрерывных (с плотностью вероятности $\varphi(x)$) приведены в табл. 3.1.

Таблица 3.1

Момент	*Случайная величина*	
	Дискретная	*Непрерывная*
Начальный	$\nu_k = \sum_{i=1}^{n} x_i^k p_i$ **(3.32)**	$\nu_k = \int_{-\infty}^{+\infty} x^k \varphi(x)dx$ **(3.33)**
Центральный	$\mu_k = \sum_{i=1}^{n} (x_i - a)^k p_i$ **(3.34)**	$\mu_k = \int_{-\infty}^{+\infty} (x-a)^k \varphi(x)dx$ **(3.35)**

Нетрудно заметить, что при $k = 1$ первый начальный момент случайной величины X есть ее математическое ожидание, т.е. $\nu_1 = M(X) = a$, при $k = 2$ второй центральный момент — дисперсия, т.е. $\mu_2 = D(X)$.

Центральные моменты μ_k могут быть выражены через начальные моменты ν_k по формулам:

$$\mu_1 = 0,$$

$$\mu_2 = \nu_2 - \nu_1^2,$$

$$\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3,$$

$$\mu_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4 \text{ и т. д.}$$

□ Например, $\mu_3 = M(X-a)^3 = M(X^3 - 3aX^2 + 3a^2X - a^3) =$ $= M(X^3) - 3aM(X^2) + 3a^2M(X) - a^3 = \nu_3 - 3\nu_1\nu_2 + 3\nu_1^2\nu_1 - \nu_1^3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3$ (при выводе учли, что $a = M(X) = \nu_1$ — неслучайная величина). ■

Выше отмечено, что математическое ожидание $M(X)$, или первый начальный момент, характеризует среднее значение или положение распределения случайной величины X на числовой оси; дисперсия $D(X)$, или второй центральный момент μ_2, — степень рассеяния распределения X относительно $M(X)$. Для более подробного описания распределения служат моменты высших порядков.

Третий центральный момент μ_3 служит для характеристики асимметрии (скошенности) распределения. Он имеет размерность куба случайной величины. Чтобы получить безразмерную величину, ее делят на σ^3, где σ — среднее квадратическое отклонение случайной величины X. Полученная величина A называется *коэффициентом асимметрии случайной величины*:

$$A = \frac{\mu_3}{\sigma^3}. \qquad (3.36)$$

Если распределение симметрично относительно математического ожидания, то коэффициент асимметрии $A = 0$.

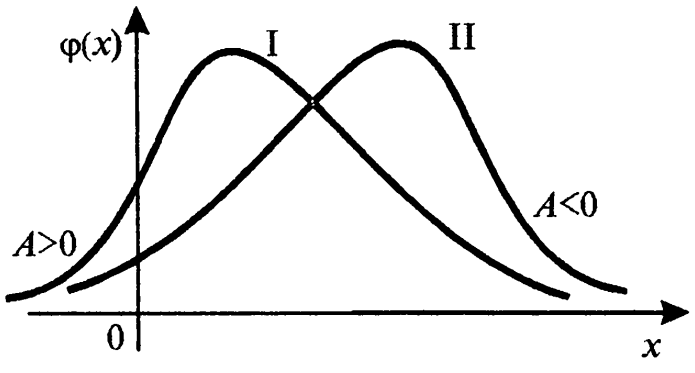


Рис. 3.17

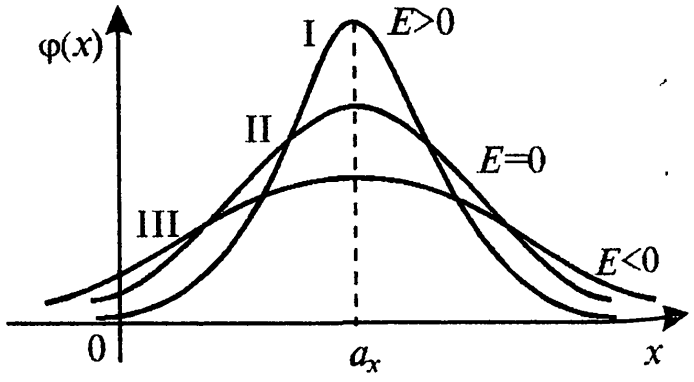


Рис. 3.18

На рис. 3.17 показаны две кривые распределения: I и II. Кривая I имеет положительную (правостороннюю) асимметрию ($A > 0$), а кривая II — отрицательную (левостороннюю) ($A < 0$).

Четвертый центральный момент μ_4 служит для характеристики крутости (островершинности или плосковершинности) распределения.

Эксцессом (или *коэффициентом эксцесса*) случайной величины называется число

$$E = \frac{\mu_4}{\sigma^4} - 3. \qquad (3.37)$$

(Число 3 вычитается из отношения μ_4/σ^4 потому, что для наиболее часто встречающегося нормального распределения (о нем идет речь в гл. 4) отношение $\mu_4/\sigma^4 = 3$. Кривые, более островершинные, чем нормальная, обладают положительным эксцессом, более плосковершинные — отрицательным эксцессом (рис. 3.18).

▷**Пример 3.17.** Найти коэффициент асимметрии и эксцесс случайной величины, распределенной по так называемому закону Лапласа с плотностью вероятности $\varphi(x) = \frac{1}{2}e^{-|x|}$.

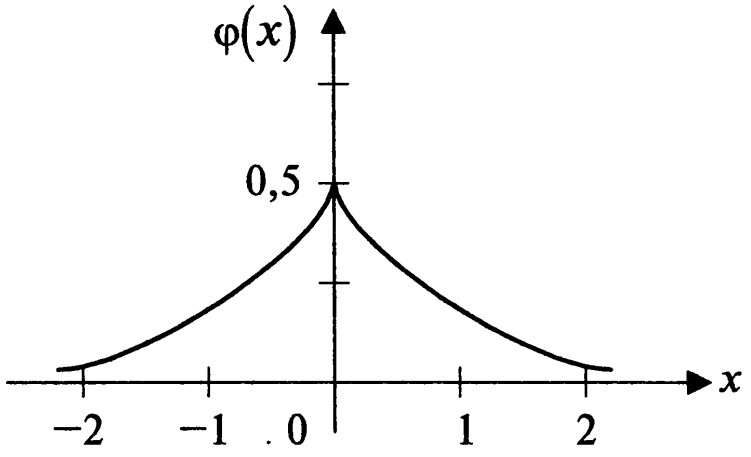


Рис. 3.19

Р е ш е н и е. Так как распределение случайной величины X симметрично относительно оси ординат, то все нечетные как начальные, так и центральные моменты равны 0, т.е. $\nu_1 = 0$, $\nu_3 = 0$, $\mu_3 = 0$ и в силу **(3.36)** коэффициент асимметрии $A = 0$.

Для нахождения эксцесса необходимо вычислить четные начальные моменты[1] ν_2 и ν_4:

$$\nu_2 = \int_{-\infty}^{+\infty} x^2\varphi(x)dx = \int_{-\infty}^{+\infty} x^2\left(\frac{1}{2}e^{-|x|}\right)dx = 2\cdot\frac{1}{2}\int_0^{+\infty} x^2e^{-x}dx = 2.$$

Следовательно,

$$D(X) = \mu_2 = \nu_2 - \nu_1^2 = 2 - 0^2 = 2 \text{ и } \sigma_x = \sqrt{D(X)} = \sqrt{2}.$$

[1] Вычисление получаемых интегралов опускаем и предлагаем его провести читателю самостоятельно.

$$\nu_4 = \int_{-\infty}^{+\infty} x^4 \varphi(x) dx = \int_{-\infty}^{+\infty} x^4 \left(\frac{1}{2} e^{-|x|} \right) dx = 2 \cdot \frac{1}{2} \int_0^{+\infty} x^4 e^{-x} dx = 24.$$

Теперь эксцесс по формуле **(3.37)**

$$E = \frac{\mu_4}{\sigma^4} - 3 = \frac{24}{\left(\sqrt{2}\right)^4} - 3 = 3.$$

Эксцесс распределения положителен, что говорит об островершинности кривой распределения $\varphi(x)$ (рис. 3.19). ▶

3.8. Решение задач

▷ **Пример 3.18.** По многолетним статистическим данным известно, что вероятность рождения мальчика равна 0,515. Составить закон распределения случайной величины X — числа мальчиков в семье из 4 детей. Найти математическое ожидание и дисперсию этой случайной величины.

Р е ш е н и е. Число мальчиков в семье из $n = 4$ представляет случайную величину X с множеством значений $X = m = 0, 1, 2, 3, 4$, вероятности которых определяются по формуле Бернулли:

$$P(X = m) = C_n^m p^m q^{n-m}, \text{ где } q = 1 - p.$$

В нашем случае $n = 4$, $p = 0{,}515$, $q = 1 - p = 0{,}485$.

Вычислим

$$P(X = 0) = C_4^0 \cdot 0{,}515^0 \cdot 0{,}485^4 = 0{,}055;$$

$$P(X = 1) = C_4^1 \cdot 0{,}515^1 \cdot 0{,}485^3 = 0{,}235;$$

$$P(X = 2) = C_4^2 \cdot 0{,}515^2 \cdot 0{,}485^2 = 0{,}375;$$

$$P(X = 3) = C_4^3 \cdot 0{,}515^3 \cdot 0{,}485^1 = 0{,}265;$$

$$P(X = 4) = C_4^4 \cdot 0{,}515^4 \cdot 0{,}485^0 = 0{,}070.$$

(Здесь учтено, что $C_4^0 = 1$, $C_4^1 = 4$, $C_4^2 = \frac{4 \cdot 3}{1 \cdot 2} = 6$, $C_4^3 = C_4^1 = 4$, $C_4^4 = 1$).

Ряд распределения имеет вид

$X=m$:	x_i	0	1	2	3	4
	p_i	0,055	0,235	0,375	0,265	0,070

Убеждаемся, что $\sum_{i=1}^{5} p_i = 0{,}055 + 0{,}235 + \ldots + 0{,}070 = 1.$

Математическое ожидание $M(X)$ и дисперсию $D(X)$ можно найти, как обычно, по формулам (**3.3**) и (**3.11**). Но в данном случае, учитывая, что закон распределения случайной величины X биномиальный (о нем см. § 4.1), можно воспользоваться простыми формулами (**4.2**) и (**4.3**):

$$M(X) = np = 4 \cdot 0{,}515 = 2{,}06,$$

$$D(X) = npq = 4 \cdot 0{,}515 \cdot 0{,}485 = 0{,}999. \blacktriangleright$$

▷ **Пример 3.19.** Радист вызывает корреспондента, причем каждый последующий вызов производится лишь в том случае, если предыдущий вызов не принят. Вероятность того, что корреспондент примет вызов, равна 0,4. Составить закон ра4спределения числа вызовов, если: а) число вызовов не более 5; б) число вызовов не ограничено.

Найти математическое ожидание и дисперсию этой случайной величины.

Р е ш е н и е: а) Случайная величина X — число вызовов корреспондента — может принимать значения 1, 2, 3, 4, 5. Обозначим событие A_i — i-й вызов принят (i = 1, 2, 3, 4, 5). Тогда вероятность того, что первый вызов принят, $P(X=1)=P(A_1)=0{,}4$.

Второй вызов состоится лишь при условии, что первый вызов не принят, т.е.

$$P(X=2) = P(\overline{A}_1 A_2) = P(\overline{A}_1)P(A_2) = (1-0{,}4)\cdot 0{,}4 = 0{,}24.$$

Аналогично

$$P(X=3) = P(\overline{A}_1 \overline{A}_2 A_3) = P(\overline{A}_1)P(\overline{A}_2)P(A_3) = 0{,}6^2 \cdot 0{,}4 = 0{,}144;$$

$$P(X=4) = P(\overline{A}_1 \overline{A}_2 \overline{A}_3 A_4) = P(\overline{A}_1)P(\overline{A}_2)P(\overline{A}_3)P(A_4) =$$

$$= 0{,}6^3 \cdot 0{,}4 = 0{,}0864.$$

Пятый вызов при любом исходе (будет принят, не принят) — последний. Поэтому

$$P(X=5)=P\left(\overline{A}_1\overline{A}_2\overline{A}_3\overline{A}_4\right)=P\left(\overline{A}_1\right)P\left(\overline{A}_2\right)P\left(\overline{A}_3\right)P\left(\overline{A}_4\right)=0{,}6^4=0{,}1296.$$

(Вероятность $P(X=5)$ можно найти и иначе, учитывая, что последний вызов будет или принят, или нет, т.е.

$$P(X=5)=P\left(\overline{A}_1\overline{A}_2\overline{A}_3\overline{A}_4A_5+\overline{A}_1\overline{A}_2\overline{A}_3\overline{A}_4\overline{A}_5\right)=$$

$$=0{,}6^4\cdot 0{,}4+0{,}6^4\cdot 0{,}6=0{,}6^4(0{,}4+0{,}6)=0{,}6^4=0{,}1296.)$$

Ряд распределения случайной величины X имеет вид

X:

x_i	1	2	3	4	5
p_i	0,4	0,24	0,144	0,0864	0,1296

Проверяем, что $\sum_{i=1}^{5} p_i = 0{,}4+0{,}24+\ldots+0{,}1296=1.$

По формуле (**3.3**) вычислим математическое ожидание:

$$a=M(X)=\sum_{i=1}^{n}x_ip_i=$$

$$=1\cdot 0{,}4+2\cdot 0{,}24+3\cdot 0{,}144+4\cdot 0{,}0864+5\cdot 0{,}1296=2{,}3056.$$

Так как $M(X)$ — нецелое число, то находить дисперсию $D(X)$ проще не по основной формуле (**3.11**), а по формуле (**3.16**), т.е. $D(X) = M(X^2) - a^2$.

Вычислим

$$M\left(X^2\right)=\sum_{i=1}^{n}x_i^2p_i=1^2\cdot 0{,}4+2^2\cdot 0{,}24+3^2\cdot 0{,}144+4^2\cdot 0{,}0864+$$

$$+5^2\cdot 0{,}1296=7{,}2784.$$

Теперь $D(X) = 7{,}2784 - 2{,}3056^2 = 1{,}9626$.

б) Так как число вызовов не ограничено, то ряд распределения случайной величины X примет вид

X:

x_i	1	2	3	4	...	n	...
p_i	0,4	0,24	0,144	0,0864	...	$0{,}6^{n-1}\cdot 0{,}4$	...

Проверяем, что

$$\sum_{i=1}^{n} p_i = 0{,}4 + 0{,}24 + \ldots + 0{,}6^{n-1} \cdot 0{,}4 + \ldots = 0{,}4\left(1 + 0{,}6 + \ldots + 0{,}6^{n-1} + \ldots\right) =$$

$$= 0{,}4 \cdot \frac{1}{1-0{,}6} = \frac{0{,}4}{0{,}4} = 1$$

(использовали формулу суммы сходящегося ($|q| < 1$) геометрического ряда: $S = \dfrac{a}{1-q}$ при $a = 1$, $q = 0{,}6$).

По формуле **(3.4)** вычислим математическое ожидание

$$M(X) = \sum_{i=1}^{\infty} x_i p_i = 1 \cdot 0{,}4 + 2 \cdot 0{,}24 + 3 \cdot 0{,}144 + \ldots + n \cdot 0{,}6^{n-1} \cdot 0{,}4 + \ldots =$$

$$= 0{,}4\left(1 + 2 \cdot 0{,}6 + 3 \cdot 0{,}6^2 + \ldots + n \cdot 0{,}6^{n-1} + \ldots\right).$$

Для вычисления суммы полученного ряда воспользуемся формулой:

$$S(x) = 1 + 2x + 3x^2 + \ldots + nx^{n-1} + \ldots = \left(x + x^2 + x^3 + \ldots + x^n + \ldots\right)' =$$

$$= \left(\frac{x}{1-x}\right)' = \frac{1}{(1-x)^2},$$

(т.е. сумма данного ряда является производной сходящегося геометрического ряда при $|q| = |x| < 1$). При $x = 0{,}6$

$$S(0{,}6) = \frac{1}{(1-0{,}6)^2} = 6{,}25, \quad \text{т.е. } M(X) = 0{,}4 \cdot 6{,}25 = 2{,}5.$$

По формуле **(3.12)** вычислим дисперсию: $D(X) = M(X^2) - a^2$.

Вначале найдем

$$M(X^2) = \sum_{i=1}^{\infty} x_i^2 p_i = 1^2 \cdot 0{,}4 + 2^2 \cdot 0{,}24 + 3^2 \cdot 0{,}144 + \ldots + n^2 \cdot 0{,}6^{n-1} \cdot 0{,}4 + \ldots =$$

$$= 0{,}4(1^2 + 2^2 \cdot 0{,}6 + 3^2 \cdot 0{,}6^2 + \ldots + n^2 \cdot 0{,}6^{n-1} + \ldots).$$

Для вычисления суммы полученного ряда рассмотрим сумму ряда $S_1(x)$ при $|x| < 1$:

$$S_1(x) = 1 + 2^2 x + 3^2 x^2 + \ldots + n^2 x^{n-1} + \ldots =$$

$$= \left(x + 2x^2 + 3x^3 + \ldots + nx^n + \ldots\right)' = \left(xS(x)\right)' = \left(\frac{x}{(1-x)^2}\right)' =$$

$$= \frac{(1-x)^2 + x \cdot 2(1-x)}{(1-x)^4} = \frac{1+x}{(1-x)^3}.$$

При $x = 0{,}6$ $S_1(0{,}6) = \dfrac{1+0{,}6}{(1-0{,}6)^3} = 25$, т.е. $M(X^2) = 0{,}4 \cdot 25 = 10$.

Теперь $D(X) = 10 - 2{,}5^2 = 3{,}75$. ▶

▷ **Пример 3.20.** Среди 10 изготовленных приборов 3 неточных. Составить закон распределения числа неточных приборов среди взятых наудачу четырех приборов. Найти математическое ожидание и дисперсию этой случайной величины.

Р е ш е н и е. Случайная величина X — число неточных приборов среди четырех отобранных — может принимать значения $i = 0, 1, 2, 3$.

Общее число способов выбора 4 приборов из 10 определяется числом сочетаний C_{10}^4. Число способов выбора четырех приборов, среди которых i неточных приборов и $4-i$ точных ($i = 0, 1, 2, 3$), по правилу произведения (см § 1.5) определится произведением числа способов выбора i неточных приборов из 3 неточных C_3^i на число способов выбора $4-i$ точных приборов из 7 точных C_7^{4-i}, т.е. $C_3^i \cdot C_7^{4-i}$. Согласно классическому определению вероятности

$$P(X = i) = \frac{C_3^i \cdot C_7^{4-i}}{C_{10}^4} (i = 0,\ 1,\ 2,\ 3).$$

Учитывая, что $C_3^0 = 1$, $C_3^1 = 3$, $C_3^2 = C_3^1 = 3$, $C_3^3 = 1$,

$C_7^4 = C_7^3 = \dfrac{7 \cdot 6 \cdot 5}{1 \cdot 2 \cdot 3} = 35$, $C_7^3 = 35$, $C_7^2 = \dfrac{7 \cdot 6}{1 \cdot 2} = 21$, $C_7^1 = 7$,

$C_{10}^4 = \dfrac{10 \cdot 9 \cdot 8 \cdot 7}{1 \cdot 2 \cdot 3 \cdot 4} = 210$, вычислим

$$P(X = 0) = \frac{35}{210} = \frac{1}{6},\quad P(X = 1) = \frac{3 \cdot 35}{210} = \frac{1}{2},$$

$$P(X=2)=\frac{3\cdot 21}{210}=\frac{3}{10},\quad P(X=3)=\frac{1\cdot 7}{210}=\frac{1}{30},$$

т.е. ряд распределения будет такой:

X:

x_i	0	1	2	3
p_i	1/6	1/2	3/10	1/30

Убеждаемся в том, что $\sum\limits_{i=1}^{4} p_i = 1/6+1/2+3/10+1/30=1.$

Математическое ожидание $M(X)$ и дисперсию $D(X)$ вычисляем по формулам **(3.3)** и **(3.16)**:

$$a = M(X) = 0\cdot\frac{1}{6}+1\cdot\frac{1}{2}+2\cdot\frac{3}{10}+3\cdot\frac{1}{30}=1{,}2,$$

$$M(X^2)=0^2\cdot\frac{1}{6}+1^2\cdot\frac{1}{2}+2^2\cdot\frac{3}{10}+3^2\cdot\frac{1}{30}=2{,}0 \text{ и}$$

$$D(X)=M(X^2)-a^2=2{,}0-1{,}2^2=0{,}56. \blacktriangleright$$

▷ **Пример 3.21.** Ряд распределения дискретной случайной величины состоит из двух неизвестных значений. Вероятность того, что случайная величина примет одно из этих значений, равна 0,8. Найти функцию распределения случайной величины, если ее математическое ожидание равно 3,2, а дисперсия 0,16.

Р е ш е н и е. Ряд распределения имеет вид

X:

x_i	x_1	x_2
p_i	0,8	0,2

,

где $p_i = 0{,}8$, а $p_2 = 1 - p_1 = 1 - 0{,}8 = 0{,}2$.

По условию

$$\begin{cases} a = M(X) = \sum\limits_{i=1}^{2} x_i p_i = 3{,}2, \\ D(X) = M(X^2) - a^2 = \sum\limits_{i=1}^{2} x_i^2 - a^2 = 0{,}16 \end{cases}$$

или $\begin{cases} 0{,}8x_1 + 0{,}2x_2 = 3{,}2, \\ 0{,}8x_1^2 + 0{,}2x_2^2 - 3{,}2^2 = 0{,}16. \end{cases}$

Решая полученную систему, находим два решения:

$$\begin{cases} x_1 = 3, \\ x_2 = 4 \end{cases} \quad \text{и} \quad \begin{cases} x_1 = 3{,}4, \\ x_2 = 2{,}4. \end{cases}$$

Аналогично примеру **3.11** записываем выражение функции распределения:

$$F(x) = \begin{cases} 0 & \text{при } x \le 3, \\ 0{,}8 & \text{при } 3 < x \le 4, \\ 1 & \text{при } x > 4 \end{cases} \text{ или } F(x) = \begin{cases} 0 & \text{при } x \le 2{,}4, \\ 0{,}2 & \text{при } 2{,}4 < x \le 3{,}4, \\ 1 & \text{при } x > 3{,}4. \end{cases} \blacktriangleright$$

▷ **Пример 3.22.** Рабочий обслуживает 4 станка. Вероятность того, что в течение часа станок не потребует внимания рабочего, для первого станка равна 0,9, для второго — 0,8, для третьего — 0,75 и для четвертого — 0,7. Составить закон распределения случайной величины X — числа станков, которые не потребуют внимания рабочего в течение часа.

Р е ш е н и е. Задача может быть решена несколькими способами.

П е р в ы й с п о с о б. Пусть $A_k\left(\overline{A}_k\right)$ — событие, состоящее в том, что k-й станок не потребует (потребует) внимания рабочего в течение часа. Тогда, очевидно:

$$P(X=0) = P\left(\overline{A}_1\overline{A}_2\overline{A}_3\overline{A}_4\right) = (1-0{,}9)(1-0{,}8)(1-0{,}75)(1-0{,}7) = 0{,}0015;$$

$$P(X=1) = P\left(A_1\overline{A}_2\overline{A}_3\overline{A}_4 + \overline{A}_1A_2\overline{A}_3\overline{A}_4 + \overline{A}_1\overline{A}_2A_3\overline{A}_4 + \overline{A}_1\overline{A}_2\overline{A}_3A_4\right) =$$

$$= 0{,}9\cdot 0{,}2\cdot 0{,}25\cdot 0{,}3 + 0{,}1\cdot 0{,}8\cdot 0{,}25\cdot 0{,}3 + 0{,}1\cdot 0{,}2\cdot 0{,}75\cdot 0{,}3 +$$

$$+ 0{,}1\cdot 0{,}2\cdot 0{,}25\cdot 0{,}7 = 0{,}0275.$$

Аналогично находим

$$P(X=2) = P\left(A_1A_2\overline{A}_3\overline{A}_4 + A_1\overline{A}_2A_3\overline{A}_4 + A_1\overline{A}_2\overline{A}_3A_4 + \overline{A}_1A_2A_3\overline{A}_4 +\right.$$

$$\left. + \overline{A}_1A_2\overline{A}_3A_4 + \overline{A}_1\overline{A}_2A_3A_4\right) = 0{,}1685;$$

$$P(X=3) = P\left(A_1A_2A_3\overline{A}_4 + A_1A_2\overline{A}_3A_4 + A_1\overline{A}_2A_3A_4 + \overline{A}_1A_2A_3A_4\right) = 0{,}4245;$$

$$P(X=4) = P(A_1A_2A_3A_4) = 0{,}378,$$

т.е. закон (ряд) распределения случайной величины X имеет вид:

X:

x_k	0	1	2	3	4
p_k	0,0015	0,0275	0,1685	0,4245	0,378

(3.38)

Второй способ состоит в том, что заданы законы (ряды) распределения альтернативных случайных величин X_k (k=1,2,3,4), выражающих число станков, не потребующих внимания рабочего в течение часа (это число для каждого станка равно 1, если этот станок не потребует внимания рабочего, и равно 0, если потребует):

X_1:

x_i	0	1
p_{i1}	0,1	0,9

X_2:

x_i	0	1
p_{i2}	0,2	0,8

X_3:

x_i	0	1
p_{i3}	0,25	0,75

X_4:

x_i	0	1
p_{i4}	0,3	0,7

Необходимо найти закон распределения суммы этих случайных величин, т.е. $X = X_1 + X_2 + X_3 + X_4$. Суммируя последовательно (см. § 3.2) $X_1 + X_2 = Z$, $X_1 + X_2 + X_3 = Z + X_3 = U$, $X_1 + X_2 + X_3 + X_4 = U + X_4 = X$, получим аналогично решению примера **3.4**:

$Z=X_1+X_2$:

z_l	0	1	2
p_l	0,02	0,26	0,72

,

$U=Z+X_3$:

u_m	0	1	2	3
p_m	0,005	0,08	0,375	0,54

и, наконец, распределение $X=U+X_4$, т.е. получили **(3.38)**.

Третий способ. Распределение X можно получить чисто механически, перемножив биномы (двучлены):

$$\varphi_4(z)=(0,1+0,9z)(0,2+0,8z)(0,25+0,75z)(0,3+0,7z), \qquad \textbf{(3.39)}$$

причем каждый из пяти полученных коэффициентов при z^k (k=0, 1, 2, 3, 4) в функции $\varphi_4(z)$ будет выражать соответствующие вероятности $P(X=k)$. Действительно, преобразовав **(3.39)**, получим

$$\varphi_4(z)=0,0015+0,0275z+0,1685z^2+0,4245z^3+0,378z^4,$$

где коэффициенты — это вероятности значений случайной величины X **(3.38)**. ▶

Функция $\varphi_n(z)=\prod_{i=1}^{n}(q_i+p_iz)$, разложение которой по степеням z дает в качестве коэффициентов вероятности значений

случайной величины X, называется *производящей функцией* для этой случайной величины. Производящая функция — полезный инструмент при описании случайной величины.

▷ **Пример 3.23.** В 1-й урне содержится 6 белых и 4 черных шара, а во 2-й −3 белых и 7 черных шаров. Из 1-й урны берут наудачу два шара и перекладывают во 2-ю урну, а затем из 2-й урны берут наудачу один шар и перекладывают в 1-ю урну. Составить законы распределения числа белых шаров в 1-й и 2-й урнах.

Р е ш е н и е. Найдем закон распределения случайной величины X — числа белых шаров в 1-й урне.

Пусть $A_i(\overline{A}_i)$ — событие, состоящее в извлечении из первой урны i-го белого (черного) шара ($i = 1, 2$), а $B(\overline{B})$ — извлечение из 2-й урны белого (черного) шара после того, как в нее из 1-й урны переложили два извлеченных шара.

В соответствии с условием число X белых шаров в 1-й урне может быть равным 4, 5, 6 или 7. Вероятность того, что в 1-й урне останется 4 белых шара, будет равна вероятности совместного осуществления трех событий: из 1-й урны извлечены первый шар — белый, второй шар — белый, из 2-й урны извлечен черный шар (после того как в нее переложили два белых шара), т.е.

$$P(X=4)=P\left(A_1A_2\overline{B}\right)=P(A_1)\cdot P_{A_1}(A_2)\cdot P_{A_1A_2}\left(\overline{B}\right)=\frac{6}{10}\cdot\frac{5}{9}\cdot\frac{7}{12}=\frac{7}{36}.$$

Рассуждая аналогично, получим

$$P(X=5)=P\left(A_1\overline{A}_2\overline{B}+\overline{A}_1A_2\overline{B}+A_1A_2e\right)=$$

$$=\frac{6}{10}\cdot\frac{4}{9}\cdot\frac{8}{12}+\frac{4}{10}\cdot\frac{6}{9}\cdot\frac{8}{12}+\frac{6}{10}\cdot\frac{5}{9}\cdot\frac{5}{12}=\frac{89}{180};$$

$$P(X=6)=P\left(\overline{A}_1\overline{A}_2\overline{B}+A_1\overline{A}_2B+\overline{A}_1A_2B\right)=$$

$$=\frac{4}{10}\cdot\frac{3}{9}\cdot\frac{9}{12}+\frac{6}{10}\cdot\frac{4}{9}\cdot\frac{4}{12}+\frac{4}{10}\cdot\frac{6}{9}\cdot\frac{4}{12}=\frac{5}{18};$$

$$P(X=7)=P\left(\overline{A}_1\overline{A}_2B\right)=\frac{4}{10}\cdot\frac{3}{9}\cdot\frac{3}{12}=\frac{1}{30}.$$

Итак, закон распределения

X:

x_i	4	5	6	7
p_i	7/36	89/180	5/18	1/30

Убеждаемся в том, что

$$\sum_{i=1}^{4} p_i = 7/36 + 89/180 + 5/18 + 1/30 = 1.$$

Распределение числа Y белых шаров во 2-й урне можно найти аналогично, но проще это сделать, если учесть, что $X+Y=9$ (при любых значениях x_i и y_j). Поэтому закон распределения случайной величины $Y = 9-X$ есть

Y:

y_j	2	3	4	5
p_j	1/30	5/18	89/180	7/36

. ▶

▷ **Пример 3.24.** Дана функция распределения случайной величины X:

$$F(x)=\begin{cases}0 & \text{при } x \le 0,\\ x^2/4 & \text{при } 0 < x \le 2,\\ 1 & \text{при } x > 2.\end{cases}$$

а) Найти плотность вероятности $\varphi(x)$; б) построить графики $\varphi(x)$ и $F(x)$; в) убедиться в том, что X — непрерывная случайная величина; г) найти вероятности $P(X=1)$, $P(X<1)$, $P(1\le X<2)$ (две последние вероятности показать на графиках $\varphi(x)$ и $F(x)$); д) вычислить математическое ожидание $M(X)$, дисперсию $D(X)$, моду $Mo(X)$ и медиану $Me(X)$.

Р е ш е н и е. а) Плотность вероятности

$$\varphi(x) = F'(x) = \begin{cases}0 & \text{при } x \le 0 \text{ и при } x > 2,\\ x/2 & \text{при } 0 < x \le 2.\end{cases}$$

б) Графики $\varphi(x)$ и $F(x)$ изображены на рис. 3.20*а* и *б*.

в) Случайная величина X — непрерывная, так как функция распределения $F(x)$ непрерывна, а ее производная — плотность вероятности $\varphi(x)$ — непрерывна во всех точках, кроме одной ($x = 2$).

г) $P(X = 1) = 0$ как вероятность отдельно взятого значения непрерывной случайной величины.

$P(X < 1)$ можно найти либо по определению функции распределения (**3.18**), либо по формуле (**3.21**) через плотность вероятности $\varphi(x)$:

$$P(X < 1) = F(1) = \frac{1^2}{4} = \frac{1}{4}$$ (ордината графика $F(1)$ — см. рис. 3.20*б* — или

$$P(X<1)=\int_{-\infty}^{1}\varphi(x)dx=\int_{-\infty}^{0}0\cdot dx+\int_{0}^{1}\frac{x}{2}dx=0+\frac{x^2}{4}\Big|_0^1=\frac{1}{4}$$

(площадь под кривой распределения $\varphi(x)$ на отрезке [0;1] — см. рис. 3.20*а*).

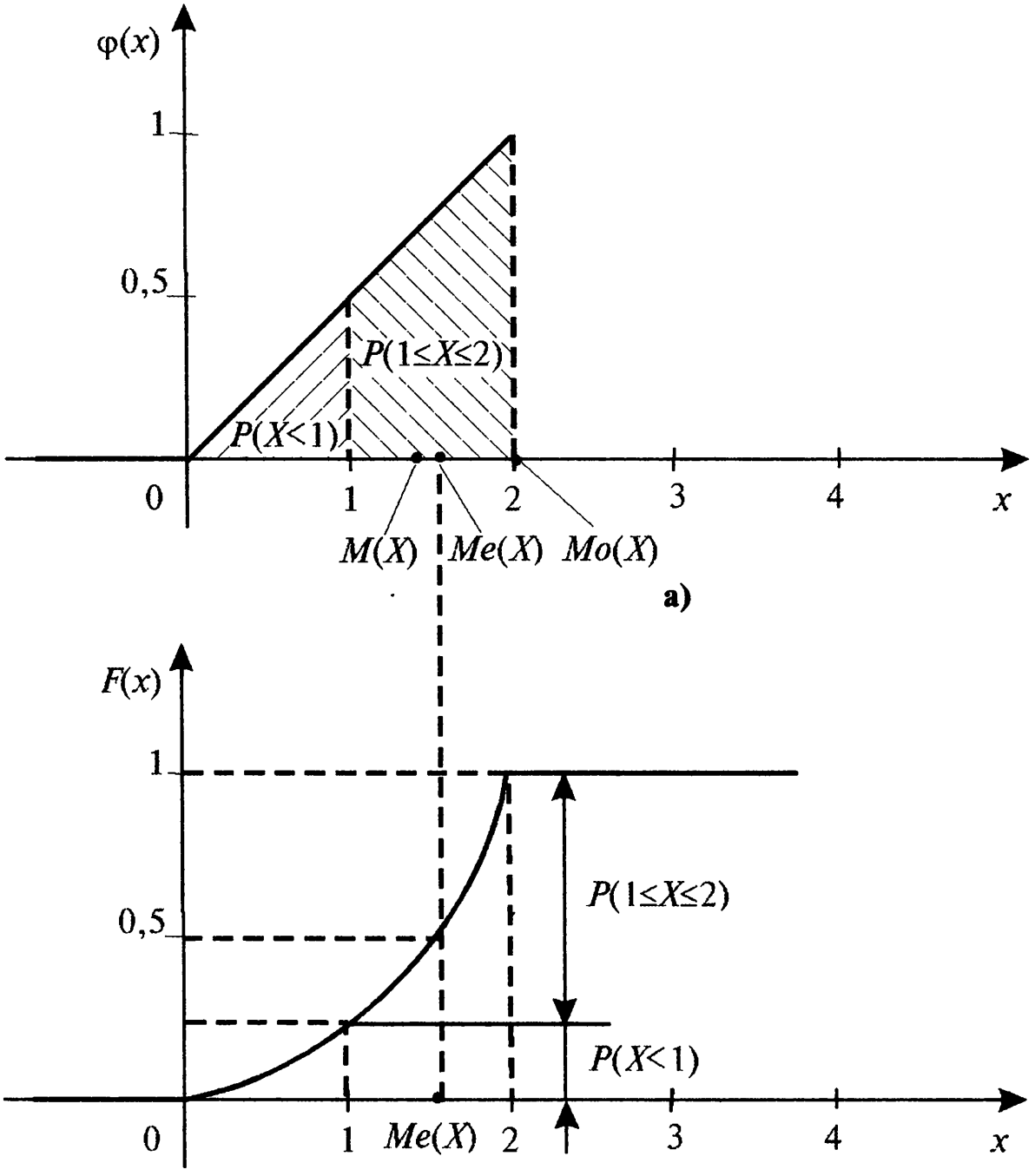


Рис. 3.20

$P(1 \le X \le 2)$ можно найти либо как приращение функции распределения по формуле (**3.20**), либо по формуле (**3.22**) через плотность вероятности $\varphi(x)$:

$$P(1 \le X \le 2) = F(2) - F(1) = \frac{2^2}{4} - \frac{1^2}{4} = \frac{3}{4}$$

(приращение ординаты графика $F(x)$ на отрезке [1;2] — рис. 3.20*б*) — или $P(1 \le X \le 2) = \int_1^2 \frac{x}{2} dx = \frac{x^2}{4}\Big|_1^2 = \frac{2^2}{4} - \frac{1^2}{4} = \frac{3}{4}$

(площадь под кривой распределения $\varphi(x)$ на отрезке [1;2] — рис. 3.20*а*).

д) По формуле (**3.25**) математическое ожидание

$$a = M(X) = \int_{-\infty}^{+\infty} x\varphi(x)dx = \int_{-\infty}^{0} 0 \cdot dx + \int_0^2 x\left(\frac{x}{2}\right)dx + \int_2^{+\infty} 0 \cdot dx =$$

$$= 0 + \frac{x^3}{6}\Big|_0^2 + 0 = \frac{1}{6} \cdot 2^3 = \frac{4}{3}.$$

Если представить распределение случайной величины X в виде единичной массы, распределенной по треугольнику (рис. 3.20*а*), то значение $M(X)=4/3$ означает абсциссу центра массы треугольника.

По формуле (**3.27**) дисперсия $D(X) = M(X^2) - a^2$.

Вначале найдем

$$M(X^2) = \int_{-\infty}^{+\infty} x^2\varphi(x)dx = 0 + \int_0^2 x^2\left(\frac{x}{2}\right)dx + 0 = 2.$$

Теперь $D(X) = 2 - \left(\frac{4}{3}\right)^2 = \frac{2}{9}$.

Плотность вероятности $\varphi(x)$ максимальна при $x = 2$ (см. рис. 3.20*а*), следовательно, $Mo(X) = 2$.

Медиану $Me(X) = b$ найдем из условия $F(b) = \frac{1}{2}$, т.е. $\frac{b^2}{4} = \frac{1}{2}$, откуда $b = Me(X) = \sqrt{2}$, или через плотность вероятности

$$\int_{-\infty}^{b} \varphi(x)dx = \frac{1}{2}, \text{ т.е. } 0 + \int_0^b \frac{x}{2} dx = \frac{x^2}{4}\Big|_0^b = \frac{b^2}{4} = \frac{1}{2},$$

откуда $b = Me(X) = \sqrt{2}$. ▶

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 рас-

пределения 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 $|X-2|\le 1$.

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: $\varphi(x)=\dfrac{A}{1+x^2}$. 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 $F(x)$.

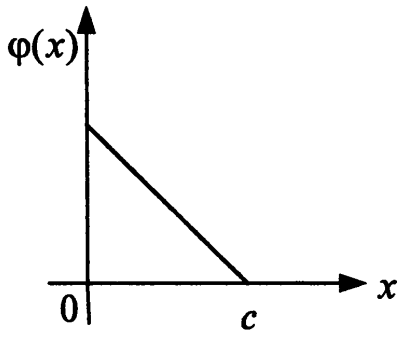


Рис. 3.21

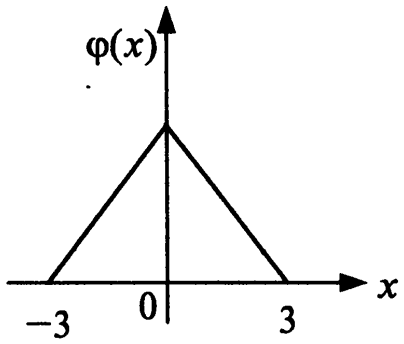


Рис. 3.22

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*

$$P(X = m) = C_n^m p^m q^{n-m}, \qquad \textbf{(4.1)}$$

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,*

$$M(X)=np, \qquad \textbf{(4.2)}$$

а ее дисперсия

$$D(X)=npq. \qquad \textbf{(4.3)}$$

□ 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,..., 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,..., 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:

$$M(X) = M(X_1+\ldots+X_k+\ldots+X_n) = \underbrace{p+\ldots+p}_{n \text{ раз}} = np,$$

$$D(X) = D(X_1+\ldots+X_k+\ldots+X_n) = \underbrace{pq+\ldots+pq}_{n \text{ раз}} = npq,$$

(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 $M(X)=np$, вы-

раженное в стандартных отклонениях $\sigma_x = \sqrt{D(X)} = \sqrt{npq}$. Аргумент $x = \dfrac{\Delta\sqrt{n}}{\sqrt{pq}} = \dfrac{\Delta}{\sqrt{pq/n}}$ 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 $\sigma\left(\dfrac{m}{n}\right) = \sqrt{D\left(\dfrac{m}{n}\right)} = \sqrt{\dfrac{pq}{n}}$.

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**):

$$M(X)=np=4\cdot 0{,}4=1{,}6,\ D(X)=npq=4\cdot 0{,}4\cdot 0{,}6=0{,}96.$$

З а м е ч а н и е. 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.

Р е ш е н и е. 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

О п р е д е л е н и е. *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

$$\sum_{i=1}^{\infty} p_i = e^{-\lambda} + \lambda e^{-\lambda} + \frac{\lambda^2 e^{-\lambda}}{2!} + \ldots + \frac{\lambda^m e^{-\lambda}}{m!} + \ldots =$$

$$= e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \ldots + \frac{\lambda^m}{m!} + \ldots\right) = e^{-\lambda} \cdot e^{\lambda} = 1$$

(учтено, что в скобках записано разложение в ряд функции e^x при $x = \lambda$).

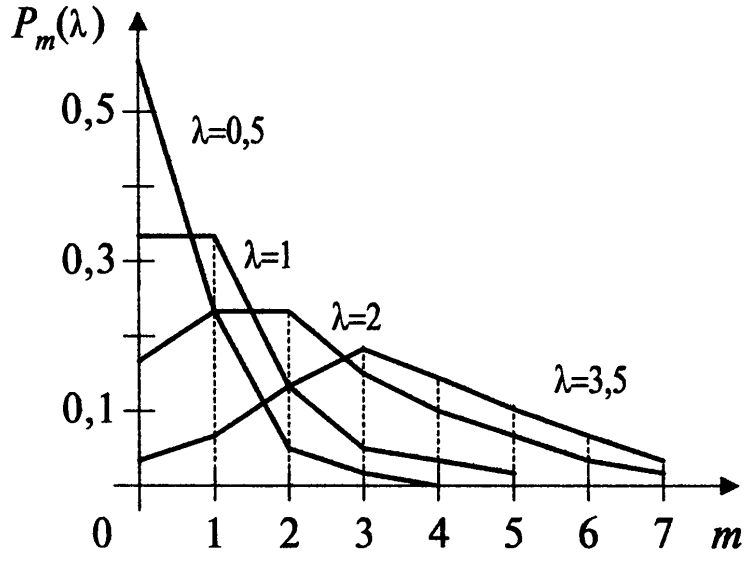


Рис. 4.1

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.

$$M(X) = \lambda, \qquad \textbf{(4.9)}$$

$$D(X) = \lambda. \qquad \textbf{(4.10)}$$

□ Найдем математическое ожидание случайной величины X:

$$a = M(X) = \sum_{i=1}^{\infty} x_i p_i = \sum_{m=0}^{\infty} m \frac{\lambda^m e^{-\lambda}}{m!} = \sum_{m=1}^{\infty} \frac{\lambda^m e^{-\lambda}}{(m-1)!} =$$

$$= \lambda e^{-\lambda} \sum_{m=1}^{\infty} \frac{\lambda^{m-1}}{(m-1)!} = \lambda e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \ldots\right) = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

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

$$P(X=m) = pq^{m-1}, \qquad (4.11)$$

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

---

[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*.

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**)

$$M(X)=\frac{1}{p}=\frac{1}{0{,}1}=10,\ D(X)=\frac{q}{p^2}=\frac{0{,}9}{0{,}1^2}=90. \blacktriangleright$$

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*

$$P(X=m)=\frac{C_M^m C_{N-M}^{n-m}}{C_N^n}, \qquad (4.14)$$

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*

$$M(X)=n\frac{M}{N}, \qquad (4.15)$$

[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*

$$D(X)=n\frac{M}{N-1}\left(1-\frac{M}{N}\right)\left(1-\frac{n}{N}\right). \qquad (4.16)$$

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.

$$\varphi(x)=\begin{cases}\dfrac{1}{b-a} & \text{при } a\le x\le b,\\ 0 & \text{при } x<a,\ x>b.\end{cases} \qquad (4.17)$$

Кривая распределения $\varphi(x)$ и график функции распределения $F(x)$ случайной величины X приведены на рис. 4.2 *а*, *б*.

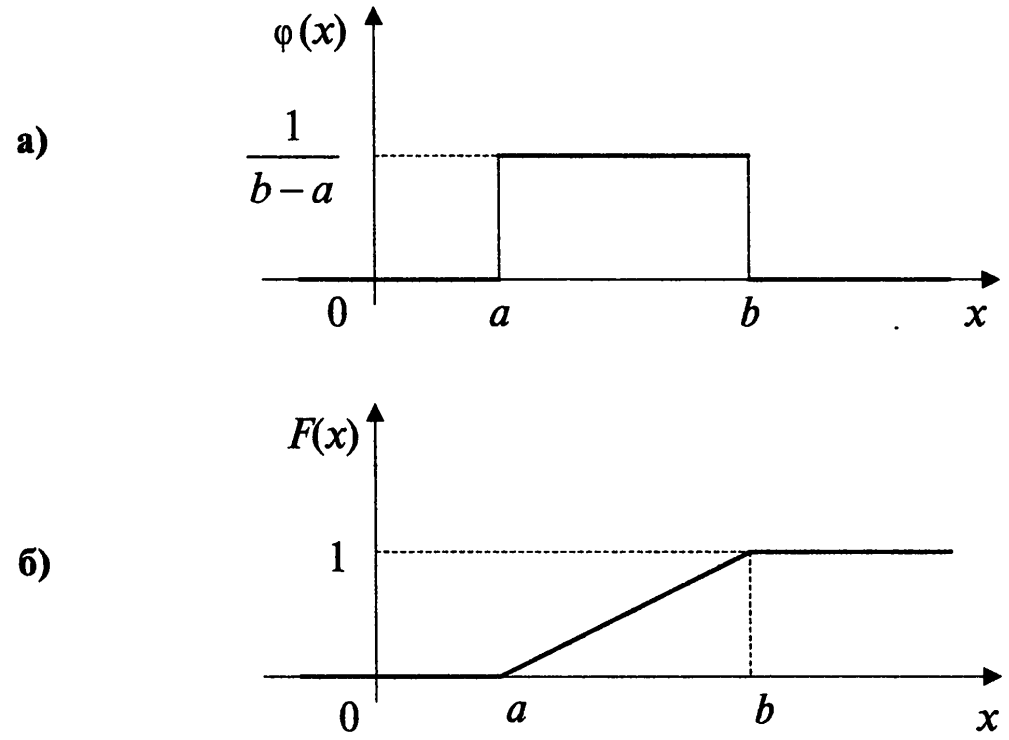


Рис. 4.2

Теорема. *Функция распределения случайной величины X, распределенной по равномерному закону, есть*

$$F(x)=\begin{cases}0 & \text{при } x\le a,\\ (x-a)\,(b-a) & \text{при } a<x\le b,\\ 1 & \text{при } x>b,\end{cases} \qquad (4.18)$$

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.

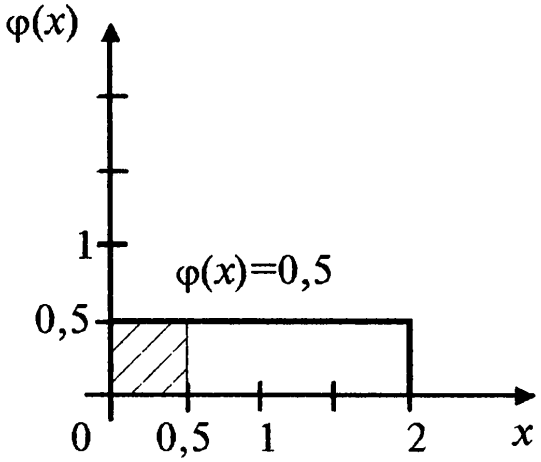


Рис. 4.3

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*

$$F(x)=\begin{cases}0 & \text{при } x<0,\\ 1-e^{-\lambda x} & \text{при } x\geq 0,\end{cases} \tag{4.22}$$

ее *математическое ожидание*

$$M(X)=\frac{1}{\lambda}, \tag{4.23}$$

а *дисперсия*

$$D(X)=\frac{1}{\lambda^2}. \tag{4.24}$$

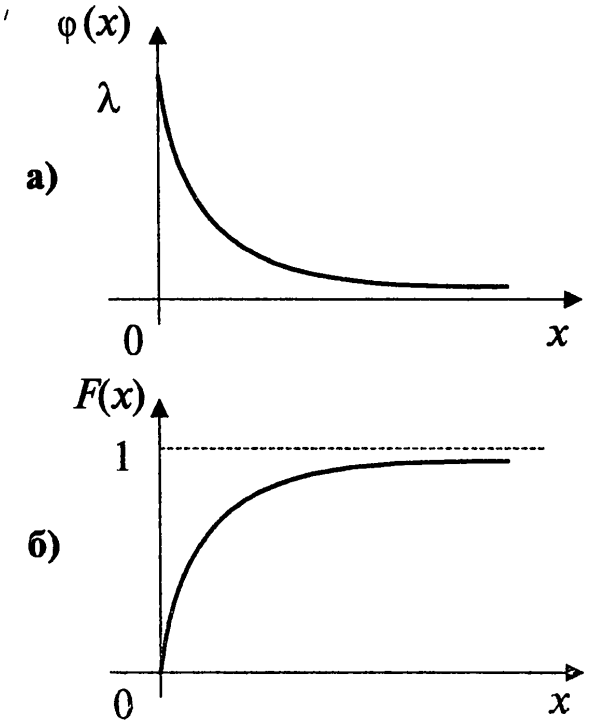


Рис. 4.4

□ При $x<0$ функция распределения $F(x) = 0$. При $x \geq a$ 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

$$F_1(t)=P_{T>\tau}(T_1<t)=\frac{P\left[(T>\tau)(T_1<t)\right]}{P(T>\tau)}. \qquad \textbf{(4.25)}$$

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.

$$P(X \geq 20) = P(20 \leq X < +\infty) = \int_{20}^{+\infty} \frac{1}{15} e^{-\frac{1}{15}x} dx,$$

но проще это сделать, используя функцию распределения:

$$P(X \geq 20) = 1 - P(X<20) = 1 - F(20) = 1 - (1 - e^{\frac{-20}{15}}) = e^{\frac{-20}{15}} = 0{,}264.$$

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:*

$$\varphi_N(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}. \qquad \textbf{(4.26)}$$

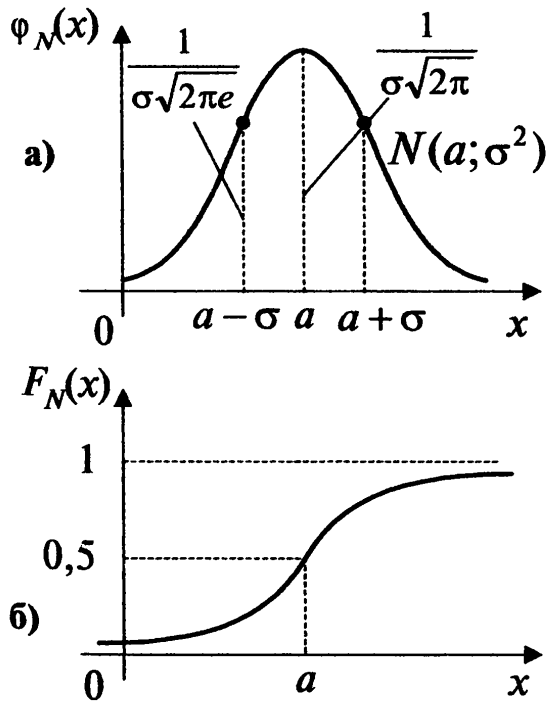


Рис. 4.5

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,

$$M(X)=\int_{-\infty}^{+\infty}\frac{1}{\sigma\sqrt{2\pi}}\left(a+\sigma\sqrt{2}t\right)e^{-t^2}\sigma\sqrt{2}\,dt=$$

$$=\frac{\sigma\sqrt{2}}{\sqrt{\pi}}\int_{-\infty}^{+\infty}te^{-t^2}dt+\frac{a}{\sqrt{\pi}}\int_{-\infty}^{+\infty}e^{-t^2}dt=0+\frac{a}{\sqrt{\pi}}\cdot\sqrt{\pi}=a$$

(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

$$D(X) = \int_{-\infty}^{+\infty} \sigma^2 2t^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-t^2} \sigma\sqrt{2}\, dt = \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{+\infty} t^2 e^{-t^2} dt = -\frac{\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{+\infty} t\, de^{-t^2}.$$

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 σ=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).

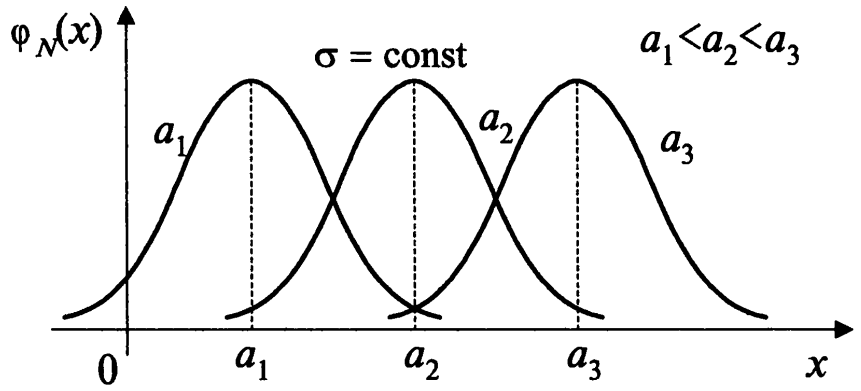


Рис. 4.6

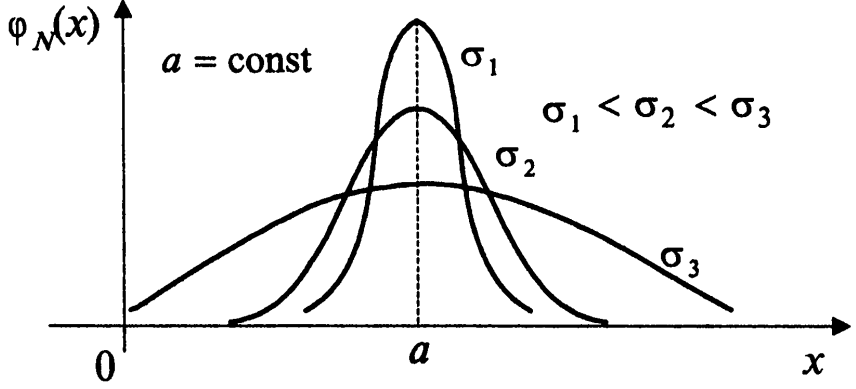


Рис. 4.7

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

$$\Phi(x) = \frac{2}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt \qquad (4.29)$$

— *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).

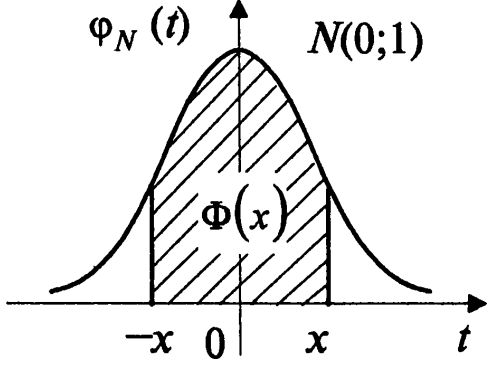


Рис. 4.8

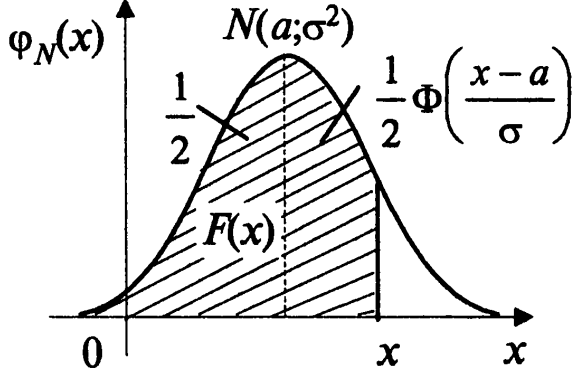


Рис. 4.9

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 $t = \frac{x-a}{\sigma}$, $x = a + t\sigma$, $dx = \sigma dt$; 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

$$\int_{-\infty}^{0} e^{-t^2/2}dt = \frac{1}{2}\int_{-\infty}^{+\infty} e^{-t^2/2}dt = \frac{1}{2}\sqrt{2}\int_{-\infty}^{+\infty} e^{-(t/\sqrt{2})^2} d\left(\frac{t}{\sqrt{2}}\right) =$$

$$= \frac{\sqrt{2}}{2}\cdot\sqrt{\pi} = \sqrt{\frac{\pi}{2}}$$

(в силу четности подынтегральной функции и того, что интеграл Эйлера—Пуассона равен $\sqrt{\pi}$).

Второй интеграл с учетом **(4.29)** составляет $\frac{1}{2}\Phi\left(\frac{x-a}{\sigma}\right)$.

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*

$$P(x_1 \le X \le x_2) = \frac{1}{2}\left[\Phi(t_2) - \Phi(t_1)\right], \qquad \textbf{(4.32)}$$

где

$$t_1 = \frac{x_1 - a}{\sigma}, \quad t_2 = \frac{x_2 - a}{\sigma}. \qquad \textbf{(4.33)}$$

□ 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

$$P(x_1 \le X \le x_2) = F(x_2) - F(x_1) =$$

$$= \left[\frac{1}{2} + \frac{1}{2}\Phi\left(\frac{x_2 - a}{\sigma}\right)\right] - \left[\frac{1}{2} + \frac{1}{2}\Phi\left(\frac{x_1 - a}{\sigma}\right)\right] = \frac{1}{2}\left[\Phi(t_2) - \Phi(t_1)\right],$$

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*

$$P(|X - a| \le \Delta) = \Phi(t), \qquad \textbf{(4.34)}$$

где

$$t = \frac{\Delta}{\sigma}. \qquad \textbf{(4.35)}$$

□ $P(|X - a| \le \Delta) = P(a - \Delta \le X \le a + \Delta)$. 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].

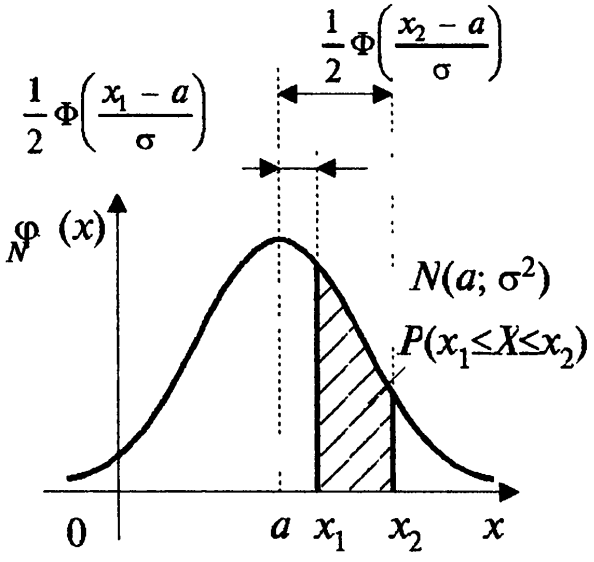


Рис. 4.10

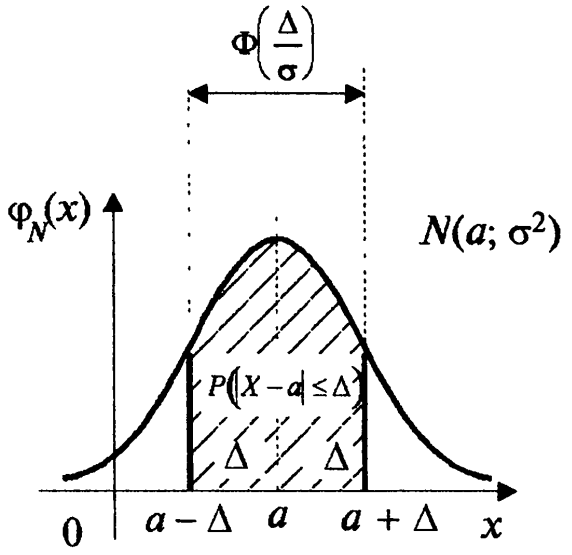


Рис. 4.11

Замечание. 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* $(a-3\sigma,\ a+3\sigma)$.

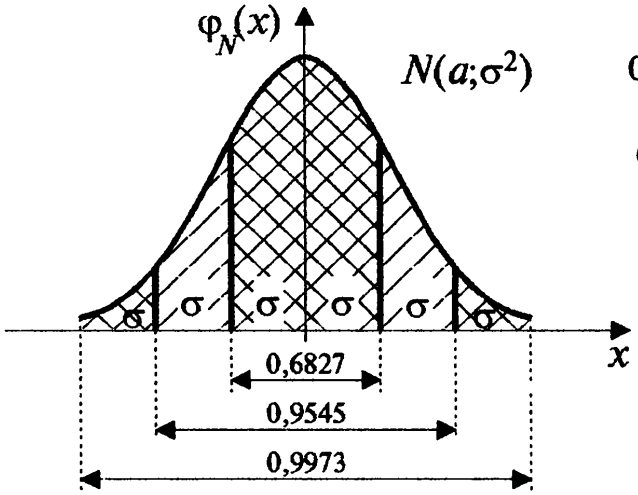


Рис. 4.12

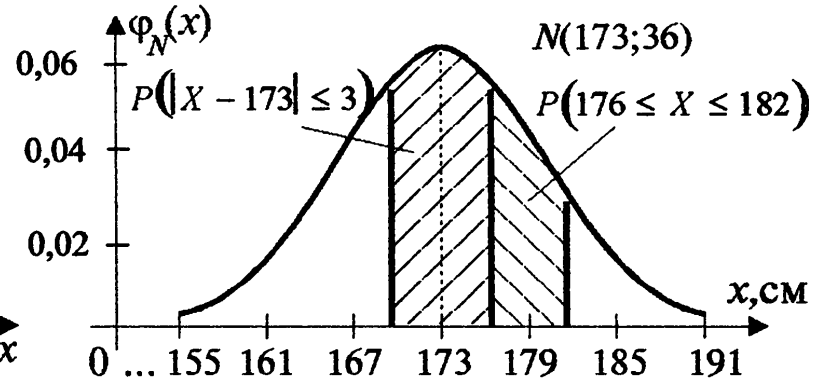


Рис. 4.13

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

$$\varphi_N(x)=\frac{1}{6\sqrt{2\pi}}e^{-\frac{(x-173)^2}{2\cdot36}};$$

$$F_N(x)=\frac{1}{6\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(x-173)^2}{2\cdot36}}dx=\frac{1}{2}+\frac{1}{2}\Phi\left(\frac{x-173}{6}\right).$$

б) Доля костюмов 4-го роста (176—182 см) в общем объеме производства определится по формуле (**4.32**) как вероятность[1]

$$P(176\le X\le182)=\frac{1}{2}[\Phi(t_2)-\Phi(t_1)]=\frac{1}{2}[\Phi(1{,}50)-\Phi(0{,}50)]=$$

$$=\frac{1}{2}(0{,}8664-0{,}3829)=0{,}2418$$

(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 сов-

падает 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 $Me(X)=a$.

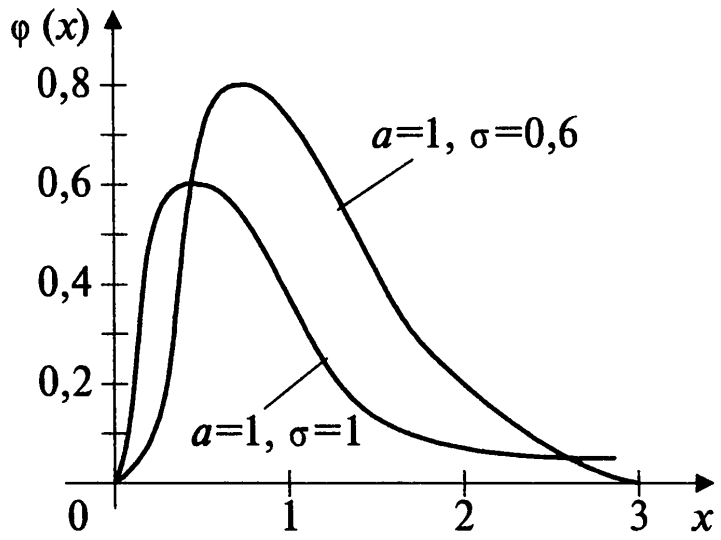


Рис. 4.14

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).

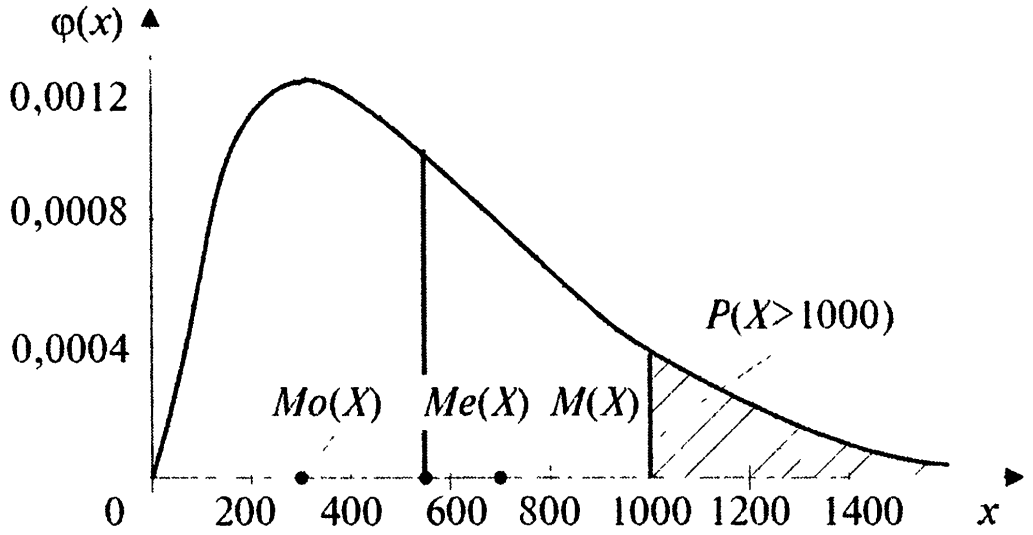


Рис. 4.15

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:

$$\varphi(x) = \begin{cases} \dfrac{1}{2^{\frac{k}{2}} \cdot \Gamma\left(\dfrac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} & \text{при } x \geq 0, \\ 0 & \text{при } x < 0, \end{cases}$$

где $\Gamma(y) = \int_0^{+\infty} e^{-t} t^{y-1} dt$ — гамма-функция Эйлера (для целых положительных значений $\Gamma(y) = (y-1)!$).

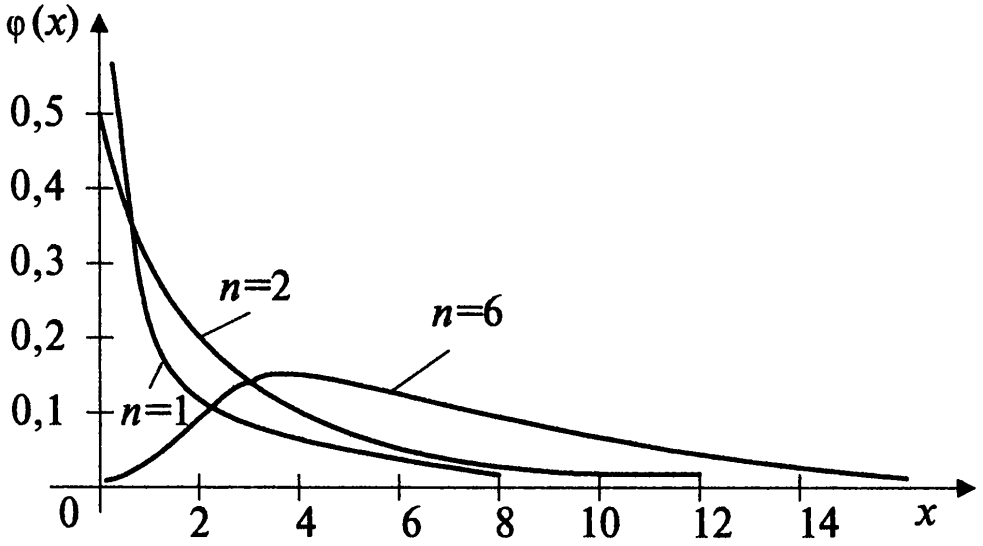


Рис. 4.16

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].

О п р е д е л е н и е. ***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:

$$\varphi(x)=\frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{\pi k}}\left(1+\frac{x^2}{n}\right)^{-\frac{k+1}{2}},$$

где $\Gamma(y)$ — гамма-функция Эйлера в точке y.

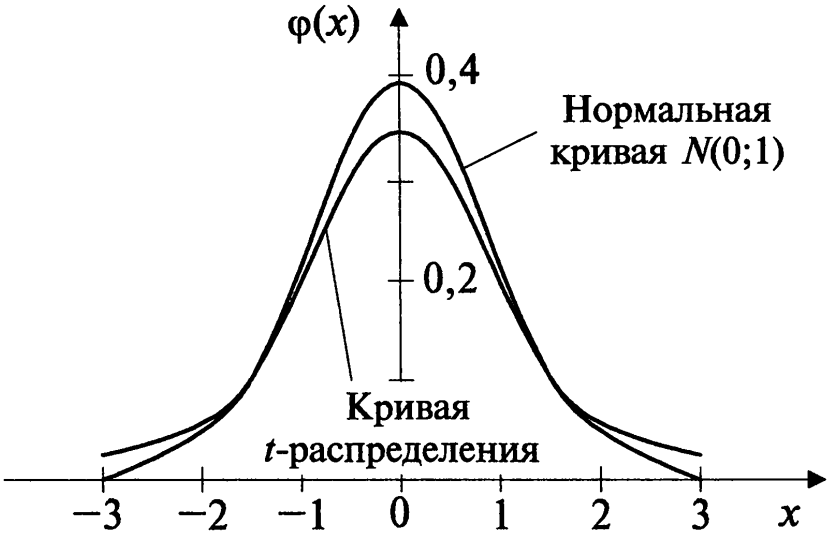


Рис. 4.17

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. $M(t)=0$, $D(t)=\dfrac{k}{k-2}$.

Распределение Фишера—Снедекора.

О п р е д е л е н и е. ***Распределением Фишера—Снедекора*** (*или* ***F-распределением***) *называется распределение случайной величины*

$$F=\frac{\frac{1}{k_1}\chi^2(k_1)}{\frac{1}{k_2}\chi^2(k_2)}, \qquad \textbf{(4.40)}$$

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:

$$\varphi(x)=\frac{\Gamma\left(\frac{k_1+k_2}{2}\right)k_1^{\frac{k_1}{2}}\,k_2^{\frac{k_2}{2}}}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)}x^{\frac{k_1}{2}-1}(k_1x+k_2)^{-\frac{k_1+k_2}{2}},$$

где $\Gamma(y)$ — гамма-функция Эйлера в точке y.

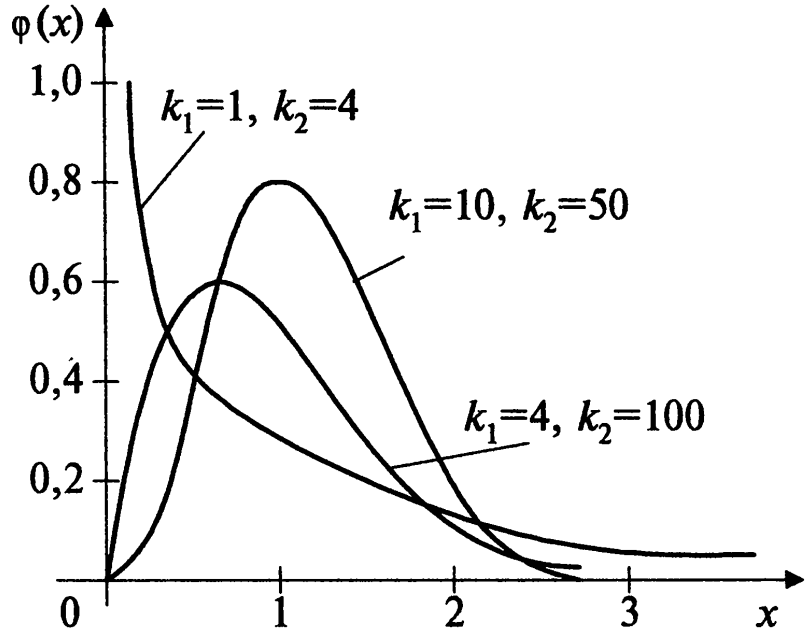


Рис. 4.18

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 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.

$$x_1 \le A, x_2 \le A, \ldots, x_k \le A; x_{k+1} > A, \ldots, x_n > A \text{ (рис. 6.1).}$$

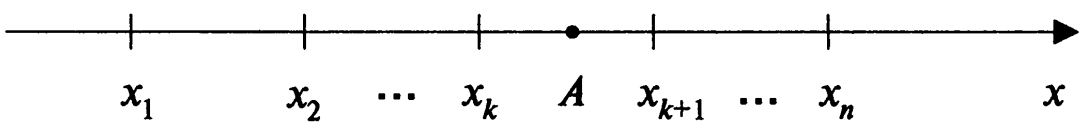


Рис. 6.1

Запишем выражение для математического ожидания $M(X)$:

$$x_1p_1+x_2p_2+\ldots+x_kp_k+x_{k+1}p_{k+1}+\ldots+x_np_n=M(X),$$

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

$$x_{k+1}p_{k+1}+\ldots+x_np_n \leq M(X). \qquad \textbf{(6.2)}$$

Заменяя в неравенстве **(6.2)** значения $x_{k+1},\ldots,x_n$ меньшим числом A, получим более сильное неравенство

$$A(p_{k+1}+\ldots+p_n) \leq M(X) \text{ или } p_{k+1}+\ldots+p_n \leq \frac{M(X)}{A}.$$

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 $P(X>A) \leq \dfrac{M(X)}{A}$. ■

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:

$$P(X \leq A) \geq 1-\frac{M(X)}{A}. \qquad \textbf{(6.3)}$$

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)** $P(X>400) \leq \dfrac{300}{400}$, т.е. вероятность того, что число вызовов превысит 400, будет не более 0,75.

б) По формуле **(6.3)** $P(X \leq 500) \geq 1-\dfrac{300}{500}=0{,}4$, 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?

Р е ш е н и е. Пусть X — размер случайно взятого вклада, а n — число всех вкладов. Тогда из условия задачи следует, что средний размер вклада $M(X)=\dfrac{2000}{n}$ (тыс. руб.). Согласно неравенству Маркова (**6.3**): $P(X \le 10) \ge 1 - \dfrac{M(X)}{10}$ или $P(X \le 10) \ge 1 - \dfrac{2000}{10n}$.

Учитывая, что $P(X \le 10) = 0{,}6$, получим $1 - \dfrac{200}{n} \le 0{,}6$, откуда $n \le 500$, т.е. число вкладчиков н е б о л е е 500. ▶

6.2. Неравенство Чебышева

Теорема. *Для любой случайной величины, имеющей математическое ожидание и дисперсию, справедливо неравенство Чебышева:*

$$P(|X - a| > \varepsilon) \le \frac{D(X)}{\varepsilon^2}, \qquad (6.4)$$

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

$$P[(X - a)^2 > \varepsilon^2] \le \frac{M(X-a)^2}{\varepsilon^2}. \qquad (6.5)$$

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 $|X - a| \le \varepsilon$ противоположны, неравенство Чебышева можно записать и в другой форме:

$$P(|X - a| \le \varepsilon) \ge 1 - \frac{D(X)}{\varepsilon^2}. \qquad (6.6)$$

Неравенство Чебышева применимо для л ю б ы х случайных величин. В форме (**6.4**) оно устанавливает в е р х н ю ю границу, а в форме (**6.6**) — н и ж н ю ю 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* $\frac{m}{n}$ *соб7ытия 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. не менее, чем 0,96. В данной задаче оценку вероятности события, найденную с помощью неравенства Маркова $(P \geq 0{,}5)$, удалось уточнить с помощью неравенства Чебышева $(P \geq 0{,}96)$. ▶

▷ **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

$$P(60 \leq m \leq 100) = P(-20 \leq m - 80 \leq 20) = P(|m - 80| \leq 20)$$

можно найти по формуле (**6.6**):

$$P(|m - 80| \leq 20) \geq 1 - \frac{2000 \cdot 0{,}04 \cdot 0{,}96}{20^2} = 1 - \frac{76{,}8}{400} = 0{,}808,$$

т.е. не менее чем 0,808.

Применяя следствие (**2.13**) интегральной теоремы Муавра—Лапласа, получим

$$P(|m - 80| \leq 20) \approx \Phi\left(\frac{20}{\sqrt{76{,}8}}\right) = \Phi(2{,}28) = 0{,}979,$$

т.е. вероятность искомого события приближенно равна 0,979.

Полученный результат $P \approx 0{,}979$ не противоречит оценке, найденной с помощью неравенства Чебышева — $P \geq 0{,}808$. Различие результатов объясняется тем, что неравенство Чебышева дает лишь нижнюю границу оценки вероятности искомого события для любой 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 определенный, 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.

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{a_1 + a_2 + \ldots + a_n}{n}\right| \le \varepsilon\right) = 1 \qquad \textbf{(6.9)}$$

или

$$\frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow[n\to\infty]{\mathcal{P}} \frac{\sum_{i=1}^{n} a_i}{n}. \qquad \textbf{(6.10)}$$

□ 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 $\left(\sum_{i=1}^{n} X_i\right)/n$ — случайная величина, а $\left(\sum_{i=1}^{n} a_i\right)/n$ — постоянное число.

Однако из (**6.9**) вовсе не следует, что это неравенство будет выполняться всегда, начиная с некоторого момента изменения $\left(\sum_{i=1}^{n} X_i\right)/n$. Так как $\left(\sum_{i=1}^{n} X_i\right)/n$ — случайная величина, то возможно, что в о т д е л ь н ы х случаях неравенство выполняться не будет. Однако с увеличением числа n вероятность неравенства $\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} a_i}{n}\right| \le \varepsilon$ 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 $\left(\sum_{i=1}^{n} X_i\right)/n$ к $\left(\sum_{i=1}^{n} a_i\right)/n$ 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 с м ы с л теоремы Чебышева. При большом числе n случайных величин $X_1, X_2,\ldots,X_n$ практически достоверно, что их средняя $X=\left(\sum_{i=1}^{n} X_i\right)/n$ — 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 $\left(\sum_{i=1}^{n} a_i\right)/n$, т.е. практически перестает быть случайной.

Следствие. *Если независимые случайные величины* $X_1, X_2,\ldots,X_n$ *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

$$\frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow[n\to\infty]{\mathcal{P}} a. \qquad \textbf{(6.15)}$$

□ 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 изме-

рений этой величины. Пусть результат каждого измерения — случайная величина X_i ($i = 1,2,..., n$). 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

$$\frac{m}{n} \xrightarrow[n\to\infty]{\mathscr{P}} p. \qquad \textbf{(6.17)}$$

□ 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

$$\frac{m}{n} \xrightarrow[n\to\infty]{\mathcal{P}} \frac{\sum_{i=1}^{n} p_i}{n}. \qquad \textbf{(6.19)}$$

□ 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 о б щ и х у с л о в и й применимости этого закона к последовательности случайных величин. Так, в теореме Маркова доказана справедливость предельного равенства **(6.15)** для з а в и с и м ы х случайных величин X_i $(i = 1,2,\ldots,n)$ 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 $(i = 1,2,\ldots,365)$ 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 принадлежит теореме Ляпунова.

Теорема Ляпунова. *Если* $X_1, X_2, \ldots, X_n$ — *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* $X_1, X_2,\ldots,X_n$ *— 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* $M\left(\left|X_i - a_i\right|^3\right) = m_i$ $(i = 1, 2,\ldots,n)$, *то закон распределения суммы* $Y_n=X_1+X_2+\ldots+X_n$ 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 $\varphi(z)$.

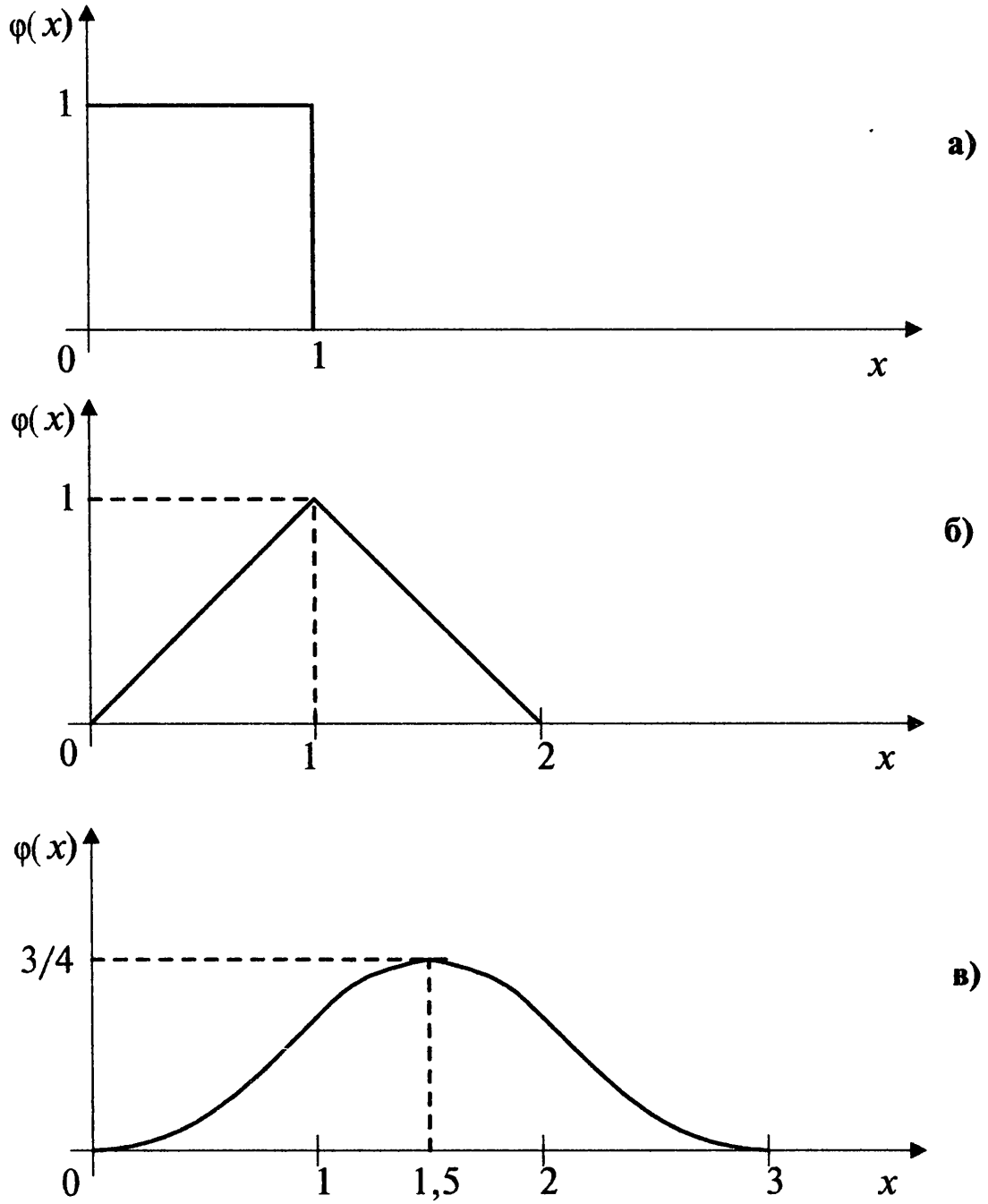


Рис. 6.2

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**):

$$P_{m,n} \approx \frac{1}{\sqrt{npq}} \varphi(z). \qquad (6.27)$$

З а м е ч а н и е. 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 с к о р о с т ь 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 измере́ний, 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?

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*.

О п р е д е л е н и е. ***Вариационным рядом*** *называется ранжированный в порядке возрастания (или убывания) ряд вариантов с соответствующими им весами (частотами или частостями)*[1].

[1] Если вариант x_i ($i = 1, 2, \ldots, n$) 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 $w_x^{\text{нак}} = n_x^{\text{нак}}/n$.

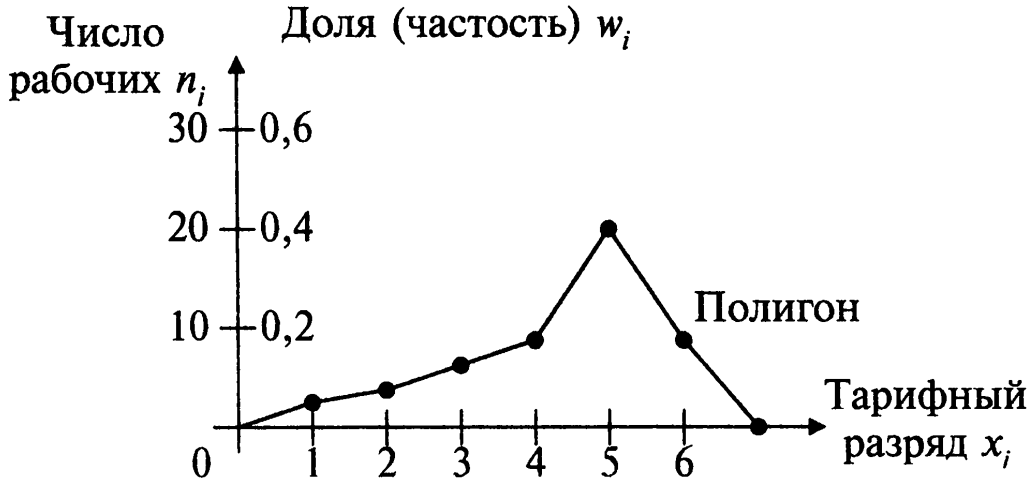
▷ **Пример 8.2.** Построить полигон (гистограмму), кумуляту и эмпирическую функцию распределения рабочих:

а) по тарифному разряду по данным табл. 8.2;

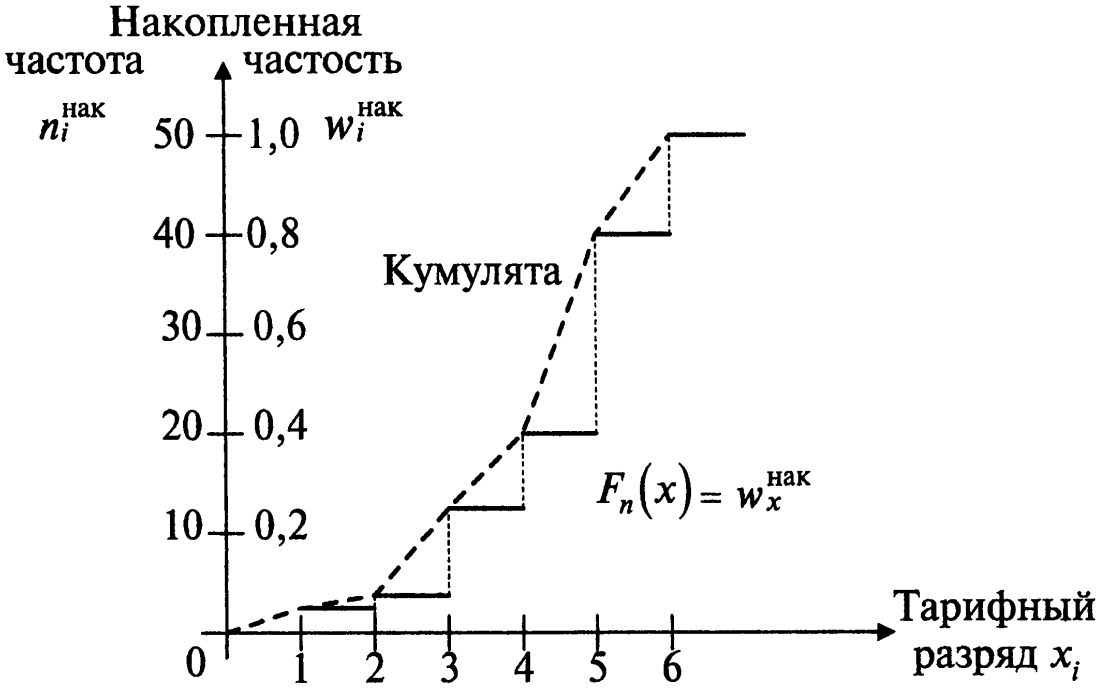
б) по выработке по данным табл. 8.1.

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 д и с к р е т н о г о 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).



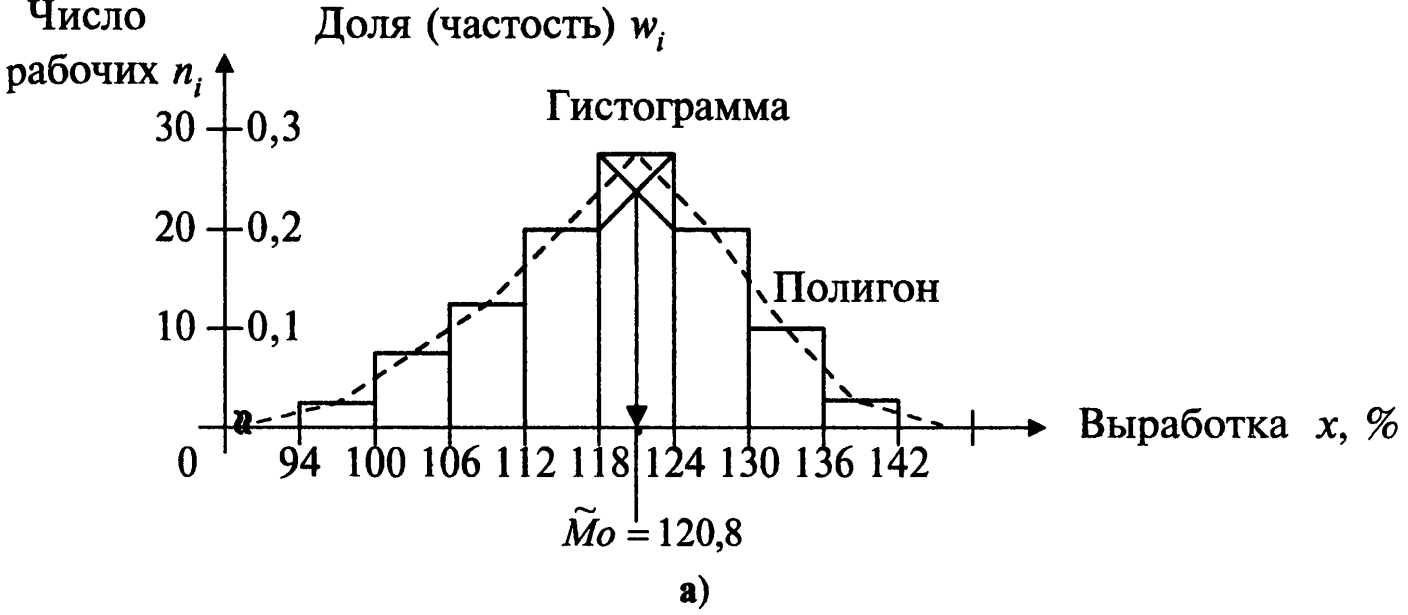
а)



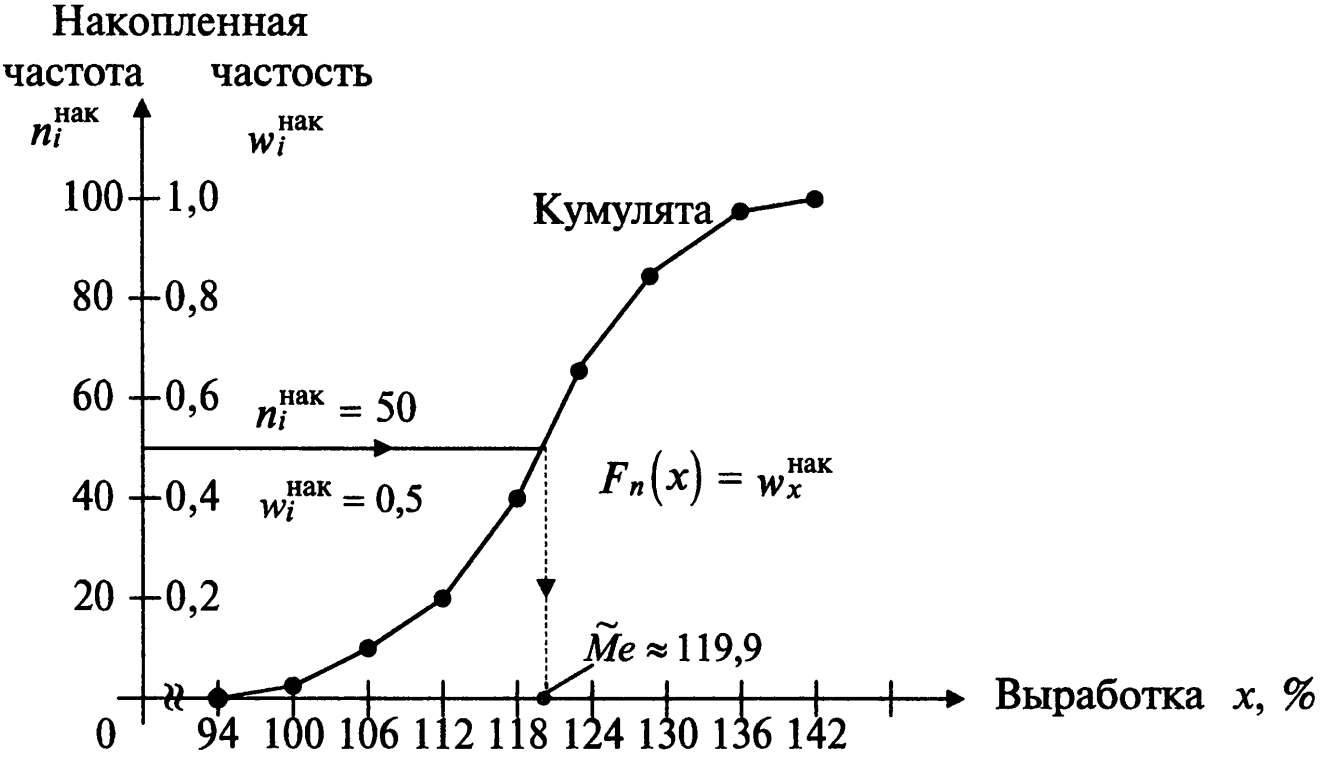
б)

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.

О п р е д е л е н и е. ***Средней арифметической*** *вариационного ряда называется сумма произведений всех вариантов на соответствующие частоты, деленная на сумму частот*:

$$\overline{x} = \frac{\sum_{i=1}^{m} x_i n_i}{n}, \qquad \textbf{(8.2)}$$

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

$$\overline{x} = \sum_{i=1}^{m} x_i w_i ,$$

где $w_i = n_i/n$ — частости вариантов или интервалов.

▷ **Пример 8.3.** Найти среднюю выработку рабочих по данным табл. 8.1.

Р е ш е н и е. По формуле **(8.2)** для интервального вариационного ряда

$$\overline{x} = \frac{97 \cdot 3 + 103 \cdot 7 + \ldots + 133 \cdot 10 + 139 \cdot 2}{100} = 119{,}2(\%),$$ где числа 97, 103,..., 133,139 — середины соответствующих интервалов. ◀

Для несгруппированного ряда все частоты n_i=1 (i=1,2,..., n), а

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \textbf{(8.3)}$$

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:*

$$\overline{kx} = k\overline{x} \text{ или } \frac{\sum_{i=1}^{m}(kx_i)n_i}{n} = k\frac{\sum_{i=1}^{m}x_i n_i}{n}.$$

3. *Если все варианты увеличить (уменьшить) на одно и то же число, то средняя арифметическая увеличится (уменьшится) на то же число:*

$$\overline{x+c} = \overline{x}+c \text{ или } \frac{\sum_{i=1}^{m}(x_i+c)n_i}{n} = \frac{\sum_{i=1}^{m}x_i n_i}{n} + c.$$

4. *Средняя арифметическая отклонений вариантов от средней арифметической равна нулю:*

$$\overline{x-\overline{x}} = 0 \text{ или } \sum_{i=1}^{m}(x_i-\overline{x})\,n_i = 0. \qquad \textbf{(8.4)}$$

□ При $c=\overline{x}$ $\overline{x-c} = \overline{x}-c = \overline{x}-\overline{x} = 0$. ■

5. *Средняя арифметическая алгебраической суммы нескольких признаков равна такой же сумме средних арифметических этих признаков:*

$$\overline{x+y} = \overline{x}+\overline{y}.$$

6. *Если ряд состоит из нескольких групп, общая средняя равна средней арифметической групповых средних, причем весами являются объемы групп:*

$$\overline{x} = \frac{\sum_{i=1}^{l}\overline{x}_i n_i}{n}, \qquad \textbf{(8.5)}$$

где $\overline{x}$ — общая средняя (средняя арифметическая всего ряда);

$\overline{x}_i$ — 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.

О п р е д е л е н и е. ***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.

Р е ш е н и е. $n = 50$ — 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.

О п р е д е л е н и е. ***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:

$$d = \frac{\sum_{i=1}^{m} |x_i - \bar{x}| n_i}{n}. \qquad \textbf{(8.6)}$$

(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 статистическим данным.

Желательно 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:*

$$s^2 = \overline{x^2} - \bar{x}^2, \tag{8.10}$$

где

$$\overline{x^2} = \frac{\sum_{i=1}^{m} x_i^2 n_i}{n}. \tag{8.11}$$

□ $$s^2 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2 n_i}{n} = \frac{\sum_{i=1}^{m} x_i^2 n_i}{n} - 2\bar{x}\frac{\sum_{i=1}^{m} x_i n_i}{n} + \bar{x}^2 \frac{\sum_{i=1}^{m} n_i}{n} =$$

$$= \overline{x^2} - 2\bar{x}\cdot\bar{x} + \bar{x}^2 = \overline{x^2} - \bar{x}^2, \text{ ибо } \sum_{i=1}^{m} n_i = n. \blacksquare$$

5. *Если ряд состоит из нескольких групп наблюдений, то общая дисперсия равна сумме средней арифметической групповых дисперсий и межгрупповой дисперсии:*

$$s^2 = \overline{s_i^2} + \delta^2, \tag{8.12}$$

где s^2 — общая дисперсия (дисперсия всего ряда);

$$\overline{s_i^2} = \frac{\sum_{i=1}^{l} s_i^2 n_i}{n} \tag{8.13}$$

— средняя арифметическая групповых дисперсий;

$$s_i^2 = \frac{\sum_{j=1}^{m}(x_j - \bar{x}_i)^2 n_j}{n_i}; \tag{8.14}$$

$$\delta^2 = \frac{\sum_{i=1}^{l}(\bar{x}_i - \bar{x})^2 n_i}{n} \tag{8.15}$$

— 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

$$\bar{u} = \left(\overline{\frac{x - c}{k}}\right) = \frac{\bar{x} - c}{k}, \qquad \textbf{(8.17)}$$

$$s_u^2 = s_{\frac{x-c}{k}}^2 = \frac{s_{x-c}^2}{k^2} = \frac{s_x^2}{k^2},$$

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

$$s_x^2 = k^2\left(\overline{u^2} - \overline{u}^2\right) = k^2\overline{u^2} - k^2\overline{u}^2 = k^2\overline{u^2} - k^2\left(\frac{\overline{x}-c}{k}\right)^2 = k^2\overline{u^2} - (\overline{x}-c)^2.$$

Теперь, заменяя в (**8.18**) и (**8.19**) $\overline{u}$ и $\overline{u^2}$ их выражениями (**8.2**) и (**8.11**) через варианты u_i, получим

$$\overline{x} = \frac{\sum_{i=1}^{m} u_i n_i}{n} \cdot k + c, \qquad (8.20)$$

$$s_x^2 = \frac{\sum_{i=1}^{m} u_i^2 n_i}{n} \cdot k^2 - (\overline{x}-c)^2, \qquad (8.21)$$

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 $\overline{x}$ 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 $\tilde{v}_k$ *k-го порядка вариационного ряда*[1] определяется по формуле:

$$\tilde{v}_k = \frac{\sum\limits_{i=1}^{m} x_i^k n_i}{n}. \qquad \textbf{(8.22)}$$

[1] См. сноску на с. 292.

Очевидно, что $\tilde{\nu}_1 = x$, т.е. средняя арифметическая является начальным моментом первого порядка вариационного ряда.

Центральный момент $\tilde{\mu}_k$ *k-го порядка вариационного ряда* определяется по формуле:

$$\tilde{\mu}_k = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^k n_i}{n}. \qquad \textbf{(8.23)}$$

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

$$\tilde{A} = \frac{\tilde{\mu}_3}{s^3} = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^3 n_i}{n s^3}. \qquad \textbf{(8.24)}$$

Если $\tilde{A} = 0$, то распределение имеет симметричную форму, т.е. варианты, равноудаленные от x, имеют одинаковую частоту. При $\tilde{A} > 0$ $\left(\tilde{A} < 0\right)$ говорят о положительной (правосторонней) или отрицательной (левосторонней) асимметрии.

Эксцессом (или *коэффициентом эксцесса*) *вариационного ряда* называется число

$$\tilde{E} = \frac{\tilde{\mu}_4}{s^4} - 3 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^4 n_i}{n s^4} - 3. \qquad \textbf{(8.25)}$$

Эксцесс является показателем «крутости» вариационного ряда по сравнению с нормальным распределением. Как отмечено выше (§ 4.7), эксцесс нормально распределенной случайной величины равен нулю.

Если $\tilde{E} > 0 \left(\tilde{E} < 0\right)$, 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

[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

1	2	3	4
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.

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 в и д ы 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 д в а с п о с о б а 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	$\overline{x}_0 = \dfrac{\sum_{i=1}^{m} x_i N_i}{N}$ **(9.1)**	$\overline{x} = \dfrac{\sum_{i=1}^{m} x_i n_i}{n}$ **(9.2)**
Дисперсия	$\sigma^2 = \dfrac{\sum_{i=1}^{m} (x_i - \overline{x}_0)^2 N_i}{N}$ **(9.3)**	$s^2 = \dfrac{\sum_{i=1}^{m} (x_i - \overline{x})^2 n_i}{n}$ **(9.4)**
Доля	$p = \dfrac{M}{N}$ **(9.5)**	$w = \dfrac{m}{n}$ **(9.6)**

Замечание. В случае бесконечной генеральной совокупности $(N = \infty)$ под генеральными средней и дисперсией понимается соответственно математическое ожидание $a = \overline{x}_0$ 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 $\bar{x}_0$, 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 о с н о в н о е у с л о в и е, которому должна удовлетворять «наилучшая» оценка.

Рассмотрим наиболее важные **свойства оценок**.

О п р е д е л е н и е. *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.*

З а м е ч а н и е. 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.***

О п р е д е л е н и е. *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 $$\tilde{\theta}_n \xrightarrow[n\to\infty]{\mathscr{P}} \theta .$$

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 $X_k (k = 1,2,...,n)$ — несмещенной $\left(M(X_k) = M(X) = \bar{x}_0\right)$, но не состоятельной оценкой генеральной средней, так как ее дисперсия $\sigma^2(X_i) = \sigma^2(X) = \sigma^2$ постоянна и не уменьшается с ростом n.

О п р е д е л е н и е. *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 р е ш а ю щ и м с в о й с т в о м, определяющим качество оценки.

---

[1] Для смещенной оценки, как нетрудно показать, $M\,(\tilde{\theta}_n - \theta)^2 = \sigma^2_{\tilde{\theta}_n} + b^2(\tilde{\theta}_n)$, где $b(\tilde{\theta}_n)$ — смещение оценки.

Эффективность оценки $\tilde{\theta}_n$ определяют отношением:

$$e = \frac{\sigma^2_{\tilde{\theta}_n^э}}{\sigma^2_{\tilde{\theta}_n}}, \qquad \textbf{(9.8)}$$

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 $\tilde{Me}$ 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. $e(\tilde{Me}) = 2/\pi = 0{,}64$ 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* $x_1, x_2,..., x_n$:

$$L(x_1, x_2,..., x_i,..., x_n; \theta) = \varphi(x_1, \theta) \cdot \varphi(x_2, \theta)...\varphi(x_i, \theta)...\varphi(x_n, \theta),$$

или $$L(x_1, x_2,..., x_i,..., x_n; \theta) = \prod_{i=1}^{n} \varphi(x_i, \theta).$$

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 $x_1, x_2,..., x_n$. И оценка $\tilde{\theta}_n$ такова, что имеющиеся у нас наблюдения $x_1, x_2,..., x_n$ 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:

$$L = (x_1, x_2,..., x_n; p) = \underbrace{pp...p}_{m}\underbrace{(1-p)(1-p)...(1-p)}_{n-m}, \text{ или}$$

$L = p^m(1-p)^{n-m}$. Тогда $\ln L = m\ln p + (n-m)\ln(1-p)$ и согласно уравнению (**9.9**)

$\frac{d\ln L}{dp} = \frac{m}{p} - \frac{n-m}{1-p}$, 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:

$$L(x_1, x_2, \ldots, x_n; a, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-a)^2}{2\sigma^2}} = \frac{1}{\sigma^n (2\pi)_2^n} e^{\frac{-\sum_{i=1}^{n}(x_i-a)^2}{2\sigma^2}}.$$

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:

$$\begin{cases} \frac{\partial \ln L}{\partial a} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - a) = 0, \\ \frac{\partial \ln L}{\partial \sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - a)^2 - \frac{n}{2\sigma^2} = 0, \end{cases}$$

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 $\tilde{\theta}_n = \dfrac{\sum\limits_{i=1}^{n} x_i}{n} = \bar{x}$, т.е. *оценка метода наименьших квадратов генеральной средней $\bar{x}_0$ есть выборочная средняя 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* случайные *величины*; *после проведения наблюдений*, когда получены их конкретные значения, — как неслучайные *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 с о с т о я т е л ь н о с т ь 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, $X_2,\ldots,X_n$ 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 $X_k=1$, $X_k=0$ ($k=1,2,\ldots,n$), а значит, зависимы случайные величины X_1, $X_2,\ldots,X_k,\ldots,X_n$.

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 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 $p(X_k=x_i)=\dfrac{N_i}{N}$ 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 $p_{X_1=x_1}(X_2=x_1)=p(X_2=x_1)=\dfrac{N_1}{N}$, 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:

$$M(X_k)=\sum_{i=1}^{m}x_i p_i=\frac{\sum_{i=1}^{m}x_i N_i}{N}=\bar{x}_0, \qquad \textbf{(9.14)}$$

$$D(X_k)=\sum_{i=1}^{m}(x_i-\bar{x}_0)^2 p_i=\frac{\sum_{i=1}^{m}(x_i-\bar{x}_0)^2 N_i}{N}=\sigma^2, \qquad \textbf{(9.15)}$$

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**):

$$\lim_{n\to\infty} P\left(\left|\bar{x} - \bar{x}_0\right| \le \varepsilon\right) = 1,$$

или

$$\bar{x} \xrightarrow[n\to\infty]{\mathcal{P}} \bar{x}_0.$$

б) **Выборка бесповторная**

В этом случае случайные величины $X_1, X_2, \ldots, X_n$ будут *зависимыми*. Рассмотрим, например, события $X_1 = x_1$ и $X_2 = x_1$.

Теперь вероятность $p_{X_1=x_1}(X_2 = x_1) = \dfrac{N_1 - 1}{N - 1}$, 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 $p(X_2 = x_1) = \dfrac{N_1}{N}$, т.е. события $X_1 = x_1$ и $X_2 = x_1$ — зависимые. Аналогично будут зависимыми любые события $X_k = x_i$ ($k = 1,2,\ldots,n$; $i = 1,2,\ldots,m$), а значит, зависимы случайные величины $X_1, X_2, \ldots, X_k, \ldots, X_n$.

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*

$$\sigma_{\bar{x}}'^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \approx \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right). \qquad (9.17)$$

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

$$s^2 = \overline{(x-x_0)^2} - (\bar{x}-\bar{x}_0)^2.$$

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:

$$M(s^2) = M\left(\frac{\sum\limits_{k=1}^{n}(X_k-\bar{x}_0)^2}{n}\right) - M(\bar{x}-\bar{x}_0)^2.$$

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. $\lim\limits_{n\to\infty}b(s^2)=0$, 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*):

$$D\left(\tilde{\theta}_n\right) \geq \frac{1}{nI(\theta)}, \qquad \textbf{(9.19)}$$

где $D\left(\tilde{\theta}_n\right)$ — 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:

$$I(\theta) = M\left[\left(\ln \varphi(X,\theta)\right)'_\theta\right]^2 = \sum_{i=1}^{m}\left[\frac{\varphi'_\theta(x_i,\theta)}{\varphi(x_i,\theta)}\right]^2 \varphi(x_i,\theta), \qquad \textbf{(9.20)}$$

а в непрерывном случае — формулой:

$$I(\theta) = M\left[\left(\ln\varphi(X,\theta)\right)'_\theta\right]^2 = \int_{-\infty}^{+\infty}\left[\frac{\varphi'_\theta(x,\theta)}{\varphi(x,\theta)}\right]^2 \varphi(x,\theta)\, d\theta. \qquad \textbf{(9.21)}$$

Неравенство информации позволяет найти тот минимум $\min D\left(\tilde{\theta}_n\right)$, который должна иметь дисперсия оценки $\sigma^2_{\tilde{\theta}_n}$, чтобы быть эффективной оценкой $\tilde{\theta}^{э}_n$, т.е. $\sigma^2_{\tilde{\theta}^{э}_n} = \min D\left(\tilde{\theta}_n\right)$.

▷ **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.
$$\min D\left(\tilde{\theta}_n\right) = \frac{1}{nI(p)} = \frac{p(1-p)}{n}.$$

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 $\bar{x}_0$ (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 $\tilde{\theta}_n^э$ 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 дис-

персии 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 одним числом, 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.

О п р е д е л е н и е. ***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).

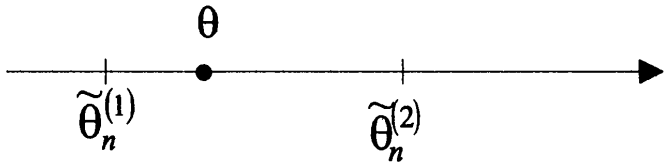


Рис. 9.1

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 с л у ч а й н ы м и в е л и ч и н а м и в отличие от оцениваемого параметра θ — величины неслучайной, поэтому правильнее говорить о том, что интервал $\left(\tilde{\theta}_n^{(1)}, \tilde{\theta}_n^{(2)}\right)$ «н а к р ы в а е т», 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 число* $\Delta > 0$ *(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 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 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	*для бесповторной выборки*
По (**9.29**)	По (**9.30**)
$\sigma_{\bar{x}} = \sqrt{\frac{87{,}48}{100}} = 0{,}935(\%)$.	$\sigma'_{\bar{x}} = \sqrt{\frac{87{,}48}{100}\left(1 - \frac{100}{1000}\right)} = 0{,}887(\%)$.

[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**):

$$P\left(\left|\bar{x}-\bar{x}_0\right|\le 1\right)=$$

$$=\Phi\left(\frac{1}{0{,}935}\right)=\Phi(1{,}07)=0{,}715.$$

$$P\left(\left|\bar{x}-\bar{x}_0\right|\le 1\right)=$$

$$=\Phi\left(\frac{1}{0{,}887}\right)=\Phi(1{,}13)=0{,}741.$$

(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 $\gamma = \Phi(t) = 0{,}9545$).

$$\Delta = 2{,}00\cdot 0{,}935 = 1{,}870(\%).$$

$$\Delta' = 2{,}00\cdot 0{,}887 = 1{,}774(\%).$$

Теперь искомый доверительный интервал определяем по (**9.27**):

$$119{,}2-1{,}870\le \bar{x}_0 \le 119{,}2+1{,}870$$

или $117{,}33\le \bar{x}_0 \le 121{,}07(\%)$.

$$119{,}2-1{,}774\le \bar{x}_0 \le 119{,}2+1{,}774$$

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**):

$$P\left(|w - p| \le 0{,}02\right) = \Phi\left(\frac{0{,}02}{0{,}0182}\right) =$$

$$= \Phi(1{,}10) = 0{,}729 \text{ (по табл. II приложений)},$$

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 сред-

ней 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)** $\sigma_w \approx \sqrt{\dfrac{w(1-w)}{n}}$ вместо точной $\sigma_w = \sqrt{\dfrac{pq}{n}} = \sqrt{\dfrac{p(1-p)}{n}}$. 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.

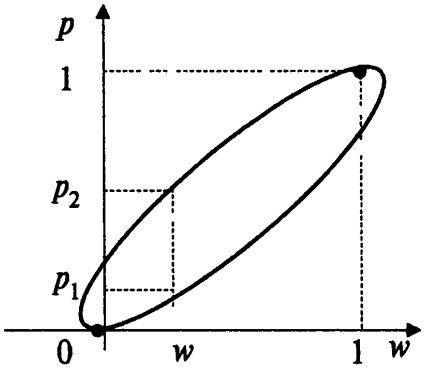


Рис. 9.2

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 у́же доверительный интервал.

Границы p_1 и p_2 доверительного интервала для p могут быть найдены из соотношения (**9.39**) по формуле:

$$p_{1,2} = \frac{1}{1+t^2/n}\left[w + \frac{t^2}{2n} \mp t\sqrt{\frac{w(1-w)}{n} + \left(\frac{t}{2n}\right)^2} \right]. \qquad (\mathbf{9.40})$$

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 $n = 50$, $w = 0{,}08$, $N = \infty$.

Р е ш е н и е. 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 $\bar{x}$ 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:

$$t=\frac{(\bar{x}-\bar{x}_0)/\dfrac{\sigma}{\sqrt{n}}}{\sqrt{\dfrac{1}{n-1}\dfrac{ns^2}{\sigma^2}}}. \qquad (9.41)$$

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 $t = \dfrac{\bar{x} - x_0}{s}\sqrt{n-1}$ 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:

$$P\left(\left|\frac{\bar{x} - \bar{x}_0}{s}\sqrt{n-1}\right| \le t_{\gamma,n-1}\right) = \theta(t, n-1) = \gamma . \qquad (\mathbf{9.42})$$

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:

$$P\left(\left|\bar{x}-\bar{x}_0\right| \le \Delta_{\text{м.в}}\right)=\theta(t,n-1)=\gamma, \qquad \textbf{(9.43)}$$

где
$$\Delta_{\text{м.в}}=\frac{t_{\gamma,n-1}s}{\sqrt{n-1}} \qquad \textbf{(9.44)}$$

— предельная ошибка малой выборки. Доверительный интервал для генеральной средней, как и ранее, находится по формуле:

$$\bar{x}-\Delta_{\text{м.в}} \le \bar{x}_0 \le \bar{x}+\Delta_{\text{м.в}}. \qquad \textbf{(9.45)}$$

▷ **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**):

$P\left(\left|\bar{x}-\bar{x}_0\right| \le 8\right)=\theta(1{,}78;\,16)=0{,}906$, ($\theta(1{,}78;\,16)$ 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 $\theta\left(t_{\gamma,k}\right)=\gamma$. 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 $\gamma = 0{,}9$.

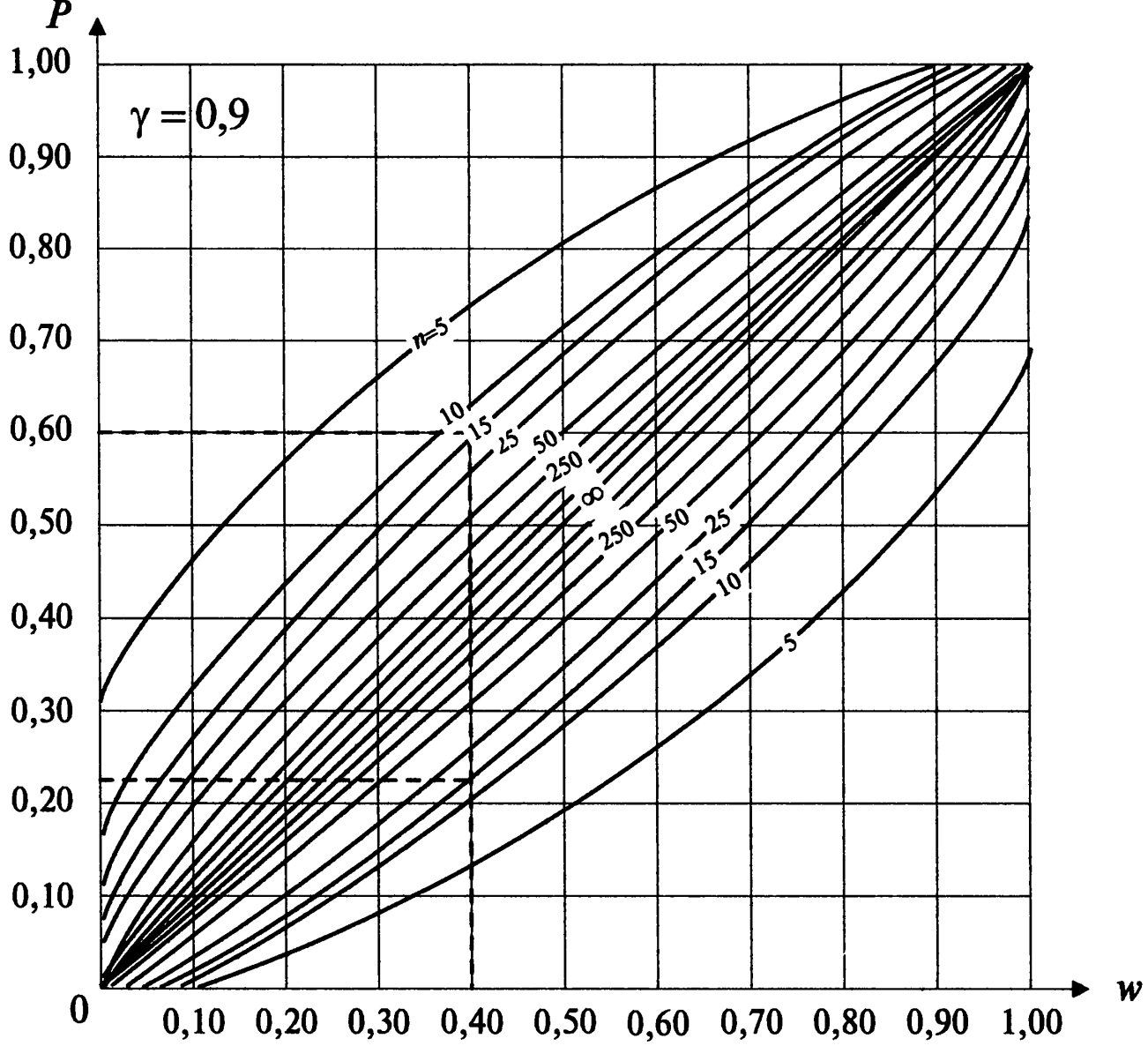


Рис. 9.3

▷ **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. $m\left(\chi^2 > \chi^2_{\alpha;k}\right) = \alpha$.

В практике выборочного наблюдения математическое ожидание 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:

$$P\left(\chi^2_1 < \frac{ns^2}{\sigma^2} < \chi^2_2\right) = \gamma \qquad \textbf{(9.46)}$$

(графически это площадь под кривой распределения между χ^2_1 и χ^2_2, см. рис. 9.4).

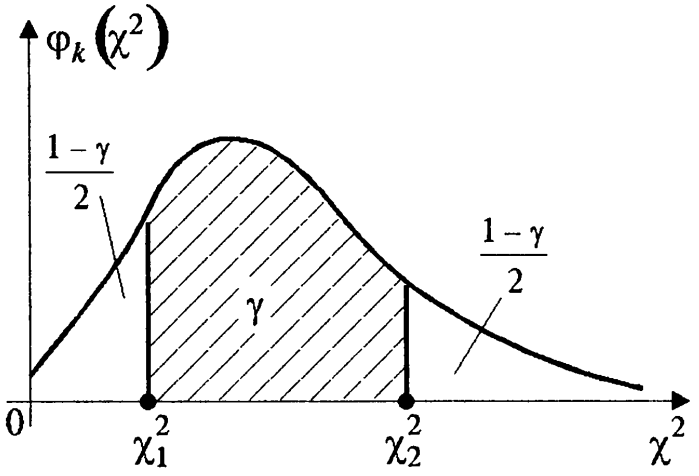


Рис. 9.4

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 $\chi^2_{\alpha;k}$, полученных из равенства $P\left(\chi^2 > \chi^2_{\alpha;k}\right) = \alpha$, 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.

$$P\left(\chi^2 > \chi_2^2\right) = \frac{1-\gamma}{2}, \qquad (9.50)$$

т.е. при $k = n - 1$ $\chi_1^2 = \chi^2_{(1+\gamma)/2;n-1}$, $\chi_2^2 = \chi^2_{(1-\gamma)/2;n-1}$.

▷ **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: $\chi_1^2 = \chi^2_{0{,}95;19} = 10{,}1$ и $\chi_2^2 = \chi^2_{0{,}05;19} = 30{,}1$. 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 $\chi^2_{\alpha;k}$ (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

$$P\left(\left|\sqrt{2\chi^2} - \sqrt{2k-1}\right| < t\right) = \Phi(t) = \gamma, \qquad (9.51)$$

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

$$\chi_1^2 = \frac{1}{2}\left(\sqrt{2 \cdot 99 - 1} - 1{,}645\right)^2 = 76{,}8, \qquad \chi_2^2 = \frac{1}{2}\left(\sqrt{2 \cdot 99 - 1} + 1{,}645\right)^2 = 122{,}9.$$

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 кото-

рых 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 дис-

персия 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.

У к а з а н и е. 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 невозмож-

ным, 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.

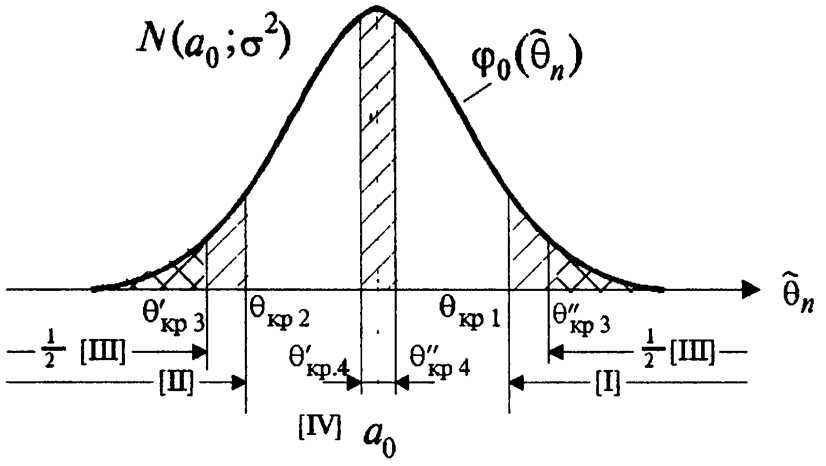


Рис. 10.1

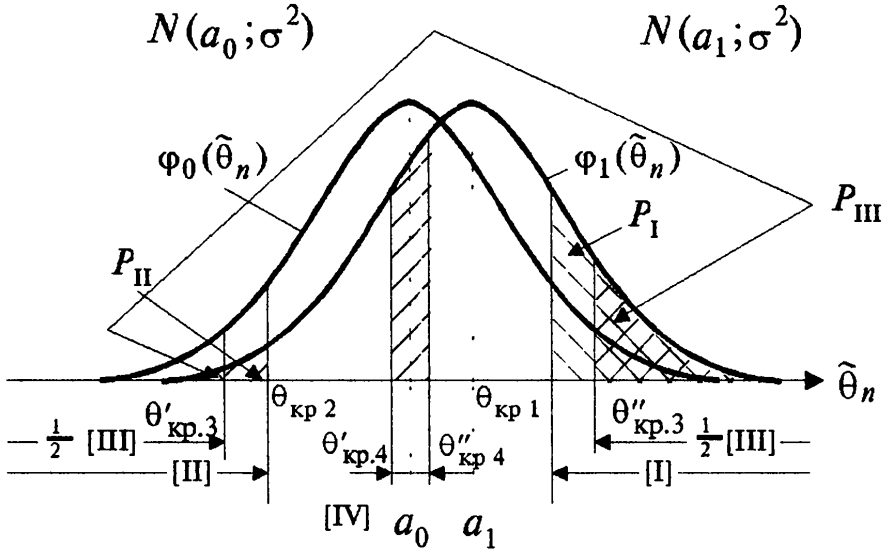


Рис. 10.2

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 распределе-

ния $\varphi_1\left(\tilde{\theta}_n\right)$ 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:

$$\begin{aligned} P\left(\tilde{\theta}_n \in W / H_0\right) &= \alpha, \\ P\left(\tilde{\theta}_n \in W / H_1\right) &= \max, \end{aligned} \qquad \textbf{(10.1)}$$

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 $(x_1, x_2, \ldots, x_n)$, имеет вид (см. § 9.3):

$$L_0(x_1, \ldots, x_n) = \varphi_0(x_1)\varphi_0(x_2)\ldots\varphi_0(x_n).$$

---

[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

$$P\left(\tilde{\theta}_n < \theta_{\text{кр}\,1}\right) = P\left(\tilde{\theta}_n > \theta_{\text{кр}.2}\right) = \frac{\alpha}{2}. \qquad \textbf{(10.4)}$$

Следует отметить, что в компьютерных статистических пакетах обычно не находятся границы критической области $\theta_{\text{кр}}$, необходимые для сравнения их с фактически наблюдаемыми значениями выборочных характеристик $\tilde{\theta}_{\text{набл.}}$ и принятия решения о справедливости гипотезы H_0. А рассчитывается точное значение уровня значимости (*p-valio*) исходя из соотношения $P(\tilde{\theta}_n > \tilde{\theta}_{\text{набл.}}) = p$. 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 генеральной 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 $N(0;1)$.

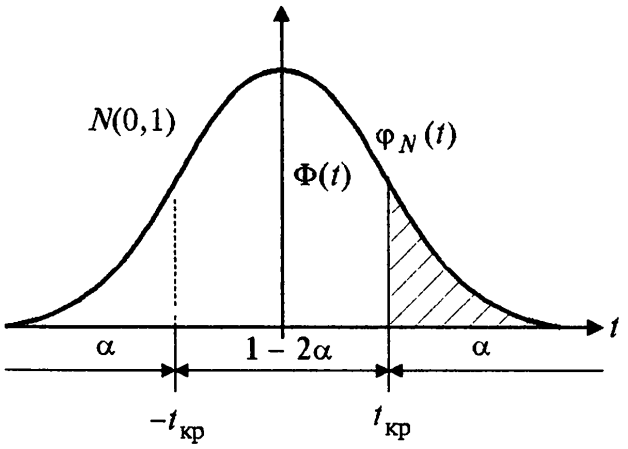


Рис. 10.3

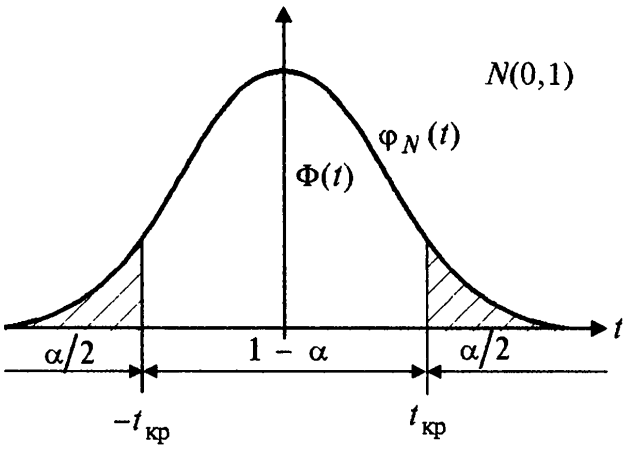


Рис. 10.4

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**)

$$\Phi\left(t_{\text{кр}}\right)=\Phi\left(t_{1-2\alpha}\right)=1-2\alpha\,, \qquad \textbf{(10.6)}$$

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**)

$$\Phi\left(t_{\text{кр}}\right)=\Phi\left(t_{1-\alpha}\right)=1-\alpha\,. \qquad \textbf{(10.7)}$$

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 $\sigma^2_{\bar{x}-\bar{y}}$ —

$$\hat{s}^2_{\bar{x}-\bar{y}} = \frac{n_1 s_x^2 + n_2 s_y^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

(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

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{n_1 s_x^2 + n_2 s_y^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \qquad \textbf{(10.8)}$$

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**)

$$t=\frac{16{,}2-13{,}9}{\sqrt{\dfrac{9\cdot 3{,}2^2+8\cdot 2{,}1^2}{8+9-2}\left(\dfrac{1}{8}+\dfrac{1}{9}\right)}}=1{,}62.$$

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, $x^*, x_1, x_2, ..., x_n$ — совокупность имеющихся наблюдений, причем x^* резко выделяется. Необходимо решить вопрос о принадлежности резко выделяющегося значения к остальным наблюдениям.

Для ряда наблюдений $x_1, x_2, ..., x_n$ 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 $|t| > t_{кр}$, и принимается, если $|t| \le t_{кр}$.

▷ **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 математиче-

1 Здесь ограничиваемся рассмотрением случая больших по объему выборок.

ским ожиданием $M(w_1 - w_2) = p - p = 0$ и дисперсией $\sigma^2_{w_1 - w_2} = \sigma^2_{w_1} + \sigma^2_{w_2} = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$. Поэтому статистика

$$t = \frac{w_1 - w_2}{\sigma_{w_1 - w_2}} = \frac{w_1 - w_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \textbf{(10.9)}$$

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)** $\hat{p} = \frac{60+69}{105+140} = \frac{129}{245} = 0{,}527$. Выборочные доли решенных задач для каждой группы $w_1 = \frac{m_1}{n_1} = \frac{60}{105} = 0{,}571$ и $w_2 = \frac{m_2}{n_2} = \frac{69}{140} = 0{,}493$. Статистика критерия по **(10.9)**

$$t=\frac{0{,}571-0{,}493}{\sqrt{0{,}527(1-0{,}527)\left(\frac{1}{105}+\frac{1}{140}\right)}}=1{,}21.$$

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 $\dfrac{(n_1-1)\hat{s}_1^2}{\sigma^2}$ и $\dfrac{(n_2-1)\hat{s}_2^2}{\sigma^2}$ имеют распределение χ^2 соответственно с $k_1=n_1-1$ и $k_2=n_2-1$ степенями свободы, а их отношение $\dfrac{\frac{1}{k_1}\chi^2(k_1)}{\frac{1}{k_2}\chi^2(k_2)}$, имеет F-распределение Фишера—Снедекора с k_1 и k_2 степенями свободы (см. § 4.9). Следовательно, случайная величина F, определяемая отношением:

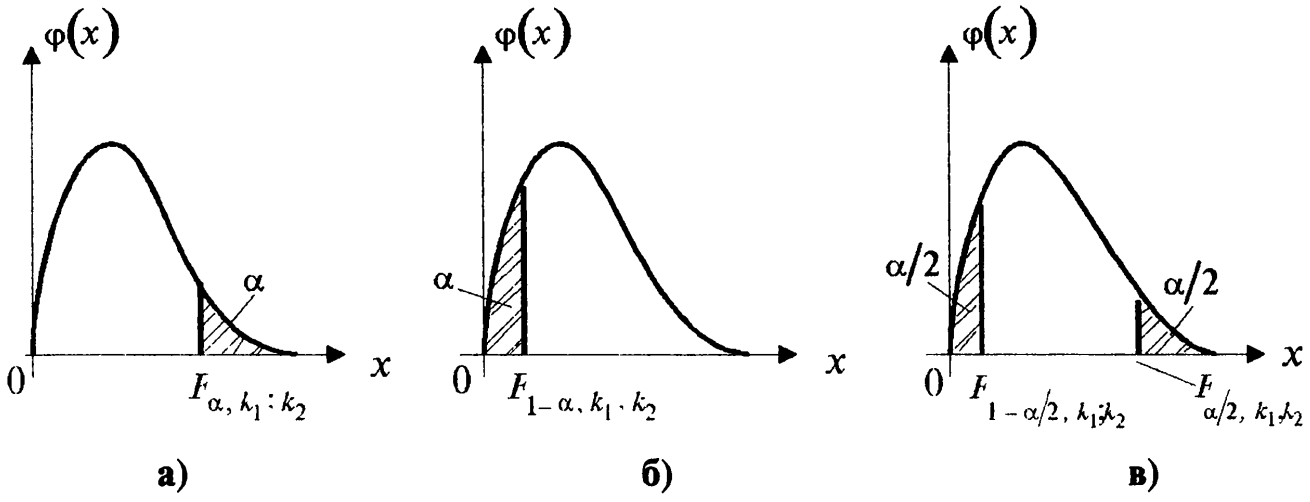
$$F = \frac{\frac{1}{n_1-1}\left[(n_1-1)\hat{s}_1^2/\sigma^2\right]}{\frac{1}{n_2-1}\left[(n_2-1)\hat{s}_2^2/\sigma^2\right]} = \frac{\hat{s}_1^2}{\hat{s}_2^2}, \qquad \textbf{(10.13)}$$

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):

$$F = \frac{\hat{s}_1^2}{\hat{s}_2^2} = \frac{\dfrac{n_1}{n_1 - 1} s_1^2}{\dfrac{n_2}{n_2 - 1} s_2^2} = \frac{(15/14) \cdot 8{,}5}{(18/17) \cdot 6{,}3} = 1{,}37 .$$

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 $n_i \geq 3 (i = 1,2,...,l)$ статистика

$$\chi^2 = \frac{\sum_{i=1}^{l} (n_i - 1) \ln\left(\overline{s^2} / \hat{s}^2\right)}{1 + \frac{1}{3(l-1)}\left(\sum_{i=1}^{l} \frac{1}{n_i - 1} - \frac{1}{n_1 + ... + n_l - l}\right)} \qquad \textbf{(10.13')}$$

(в которой

$$\hat{s}^2 = \frac{n_i s_i^2}{n_i - 1} \qquad \textbf{(10.14)}$$

— исправленная выборочная дисперсия i-й выборки,

$$\overline{s^2} = \frac{1}{n_1 + ... + n_l - l} \sum_{i=1}^{l} n_i s_i^2 \qquad \textbf{(10.15)}$$

— 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	$t=\dfrac{\bar{x}-a_0}{\sigma/\sqrt{n}}$	$a=a_1>a_0$ $a=a_1<a_0$	$\lvert t\rvert > t_{1-2\alpha}$
			$a=a_1\neq a_0$	$\lvert t\rvert > t_{1-\alpha}$
	σ^2 неизвестна	$t=\dfrac{\bar{x}-a_0}{s/\sqrt{n-1}}$	$a=a_1>a_0$ $a=a_1<a_0$	$\lvert t\rvert > t_{1-2\alpha,n-1}$
			$a=a_1\neq a_0$	$\lvert t\rvert > t_{1-\alpha,n-1}$
$\sigma^2=\sigma_0^2$	a неизвестно	$\chi^2=\dfrac{ns^2}{\sigma_0^2}$	$\sigma^2=\sigma_1^2>\sigma_0^2$	$\chi^2>\chi^2_{\alpha,n-1}$
			$\sigma^2=\sigma_1^2<\sigma_0^2$	$\chi^2<\chi^2_{1-\alpha,n-1}$
			$\sigma^2=\sigma_1^2\neq\sigma_0^2$	$\chi^2>\chi^2_{\alpha/2;n-1}$ либо $\chi^2<\chi^2_{1-\alpha/2,n-1}$
$p=p_0$	Достаточно большие n	$t=\dfrac{w-p_0}{\sqrt{p_0q_0/n}}$	$p=p_1>p_0$ $p=p_1<p_0$	$\lvert t\rvert > t_{1-2\alpha}$
			$p=p_1\neq p_0$	$\lvert t\rvert > t_{1-\alpha}$

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 θ (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, $P=\frac{1}{2}-\frac{1}{2}\theta(-0{,}42;9)\approx\frac{1}{2}(1+0{,}31)\approx 0{,}66$. ▶

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 $t=\frac{\bar{x}-a_0}{s/\sqrt{n-1}}$ 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 $1-\beta$ = 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 $20\neq\frac{16}{9}(t_{0,9;19}+t_{0,95;19})^2=\frac{16}{9}(1{,}73+2{,}09)^2=24{,}7)$, а при $n=25$ $25\approx\frac{16}{9}(t_{0,9;24}+t_{0,95;24})^2=\frac{16}{9}(1{,}71+2{,}06)^2=25{,}3$.

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. $\theta(0{,}40;\ 9)$ = 0,3, и $\theta(0{,}54;\ 9)$ = 0,4, а искомое значение $\theta(0{,}42;\ 9)\approx 0{,}31$ 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%.

Р е ш е н и е. 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%.

Р е ш е н и е. 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 $\alpha=0,1$ проверить гипотезу о том, что среднее квадратическое отклонение суточной выработки работниц равно 20 м/ч.

Р е ш е н и е. 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 $\alpha=0,1$ 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 $\chi^2 \le \chi^2_{\alpha;k}$, гипотеза H_0 не противоречит опытным данным.

З а м е ч а н и е. 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 $\bar{x} = 119{,}2(\%)$, $s^2 = 87{,}48$, $s = 9{,}35(\%)$.

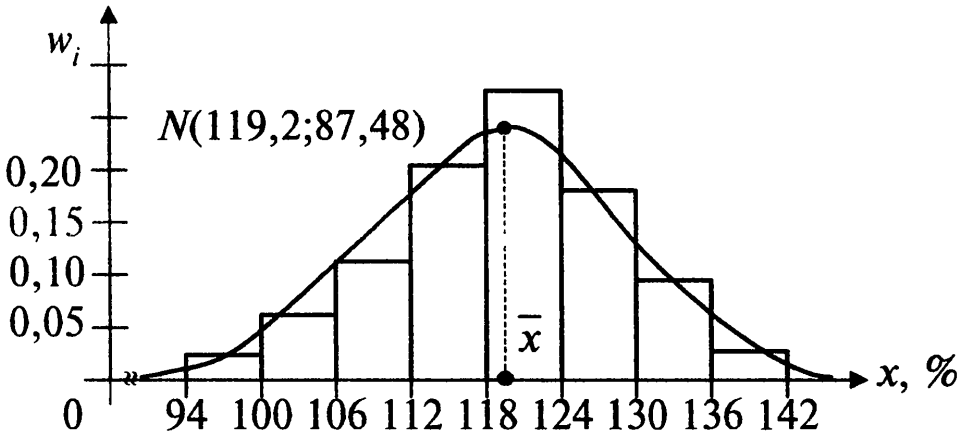


Рис. 10.6

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, $p_1(94 \le X \le 100) = \frac{1}{2}\left[\Phi\left(\frac{100-119{,}2}{9{,}35}\right) - \Phi\left(\frac{94-119{,}2}{9{,}35}\right)\right] =$

$= \frac{1}{2}[\Phi(-2{,}05) - \Phi(-2{,}69)] = \frac{1}{2}(-0{,}9596 + 0{,}9928) = 0{,}0166$ 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 $\chi^2 = 2{,}27$.

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 $\varphi_N(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\alpha)^2/2\sigma^2}$, 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 п р и б л и ж е н н ы й 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 каче-

стве 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:

$$\bar{x} = \frac{\sum_{i=1}^{m} x_i n_i}{n} = \frac{0 \cdot 15 + \ldots + 8 \cdot 1}{300} = 2{,}54;$$

$$s^2 = \overline{x^2} - \bar{x}^2 = \frac{\sum_{i=1}^{m} x_i^2 n_i}{n} - x^2 = \frac{0^2 \cdot 15 + \ldots + 8^2 \cdot 1}{300} - 2{,}54^2 \approx 2{,}39.$$

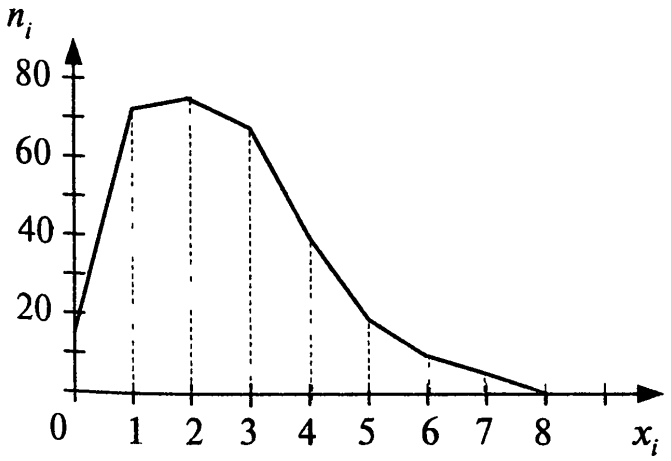


Рис. 10.6а

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 распределе-

ниями 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

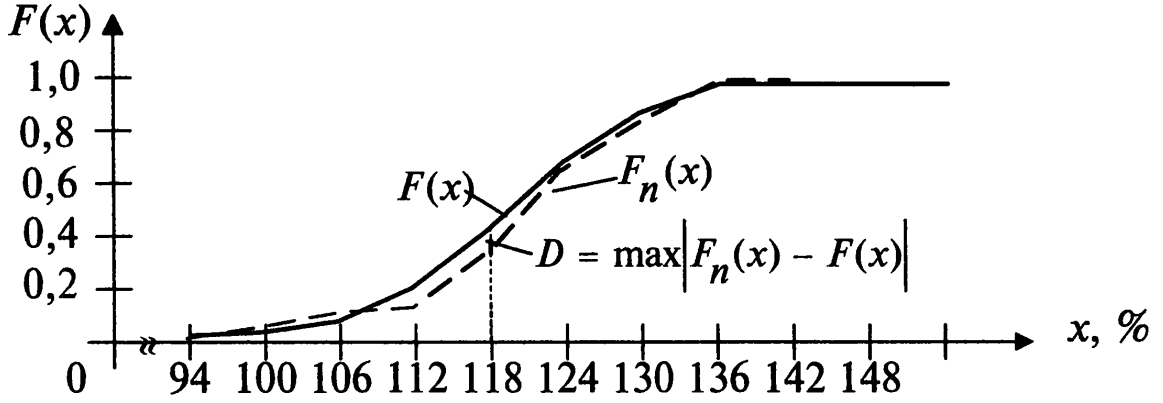


Рис. 10.7

Из рис. 10.7 следует, что

$$D = \left|F_n(118) - F(118)\right| = \left|0{,}410 - 0{,}449\right| = 0{,}039.$$

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: $n_{i_1}^{\text{нак}}$ и $n_{i_2}^{\text{нак}}$ — накопленные частоты соответственно выборок 1 и 2; $F_{n_1}(x_i) = n_{i_1}^{\text{нак}}/n_1$, $F_{n_2}(x_i) = n_{i_2}^{\text{нак}}/n_2$ — значения их эмпирических функций распределения. Результаты вычислений сведем в табл. 10.7.

Таблица 10.7

x_i	$n_{i_1}^{\text{нак}}$	$n_{i_2}^{\text{нак}}$	$F_{n_1}(x_i)$	$F_{n_2}(x_i)$	$\lvert F_{n_1}(x_i) - F_{n_2}(x_i)\rvert$
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 $\lambda' = \sqrt{\dfrac{110 \cdot 100}{110+100}} \cdot 0{,}089 = 0{,}644$. 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

$$\chi^2 = n\sum_{i=1}^{m}\sum_{j=1}^{l}\frac{(n_{ij} - n_{i*}n_{*j}/n)^2}{n_{i*}n_{*j}} = n\left(\sum_{i=1}^{m}\sum_{j=1}^{l}\frac{n_{ij}^2}{n_{i*}n_{*j}} - 1\right), \qquad \textbf{(10.22)}$$

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 $\bar{y}=185$ 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 $H_1: \bar{x}_0 \neq \bar{y}_0$; б) конкурирующая гипотеза $H_1: \bar{x}_0 < \bar{y}_0$.

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, $\bar{x} = 0{,}53$ 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 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:

$$M_x(Y) = \varphi(x) \qquad \textbf{(12.1)} \quad \text{или} \quad M_y(X) = \psi(y). \qquad \textbf{(12.2)}$$

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 яв-

---

[1] Для условного математического ожидания в литературе используется также обозначение $M(Y \mid X = x)$.

[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).

ляется 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

$$y_x = \hat{\varphi}(x, b_0, b_1, ..., b_p) \qquad \textbf{(12.3)}$$

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:

$$x_y = \hat{\psi}(y, c_0, c_1, ..., c_p), \qquad \textbf{(12.4)}$$

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] 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 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.*

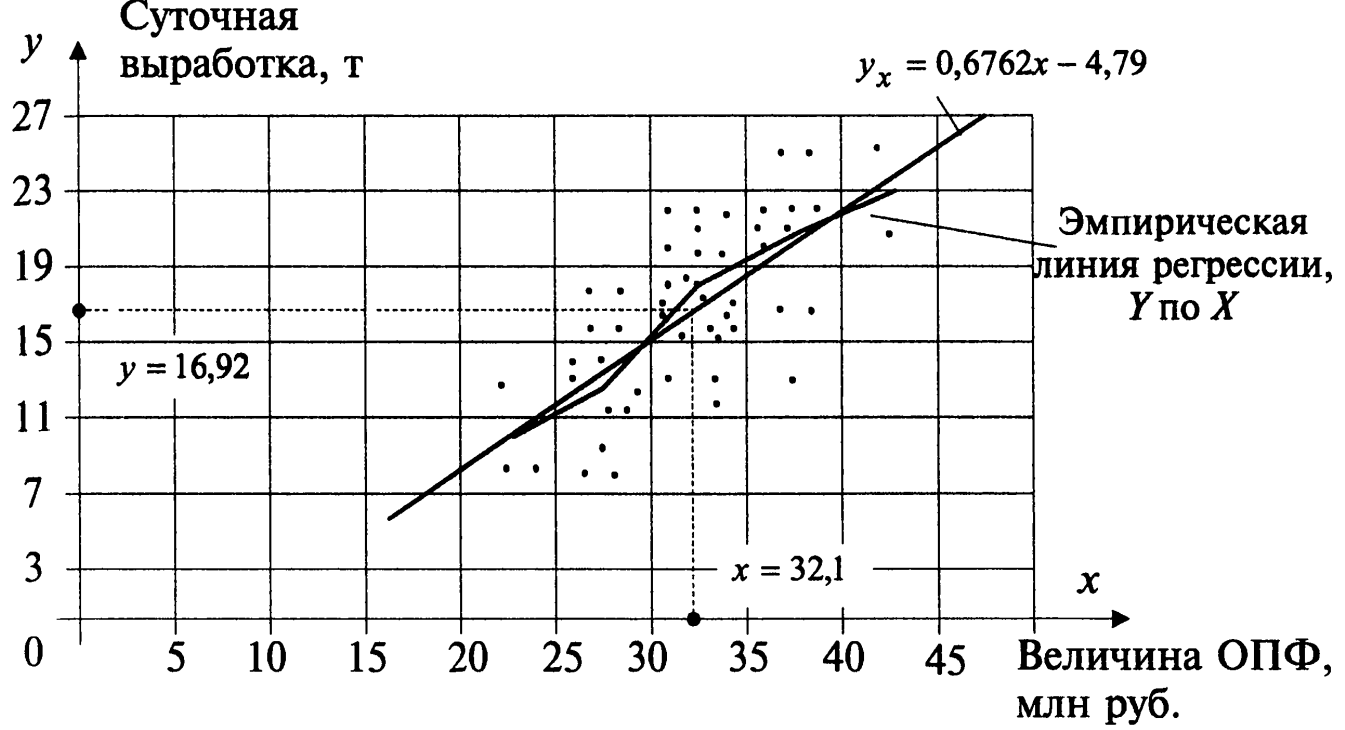


Рис. 12.1

Для каждого значения x_i ($i = 1,2,...,l$), т.е. для каждой строки корреляционной таблицы вычислим групповые средние

$$\bar{y}_i = \frac{\sum_{j=1}^{m} y_j n_{ij}}{n_i}, \qquad \textbf{(12.5)}$$

где n_{ij} — частоты пар (x_i, y_j) и $n_i = \sum_{i=1}^{m} n_{ij}$; 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

$$\bar{x}_j = \frac{\sum_{i=1}^{l} x_i n_{ij}}{n_j} \qquad \textbf{(12.6)}$$

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 л и н е й н о й корреляционной зависимости Y по X между двумя рассматриваемыми переменными, которая графически выражается тем точнее, чем больше объем выборки (число рассматриваемых предприятий) n:

$$n = \sum_{i=1}^{l} n_i = \sum_{j=1}^{m} n_j = \sum_{i=1}^{l} \sum_{j=1}^{m} n_{ij} . \qquad \textbf{(12.7)}$$

Поэтому уравнение регрессии **(12.3)** будем искать в виде:

$$y_x = b_0 + b_1 x . \qquad \textbf{(12.8)}$$

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*:

$$S = \sum_{i=1}^{l} \left(y_{x_i} - \bar{y}_i \right)^2 n_i = \sum_{i=1}^{l} \left(b_0 + b_1 x_i - \bar{y}_i \right)^2 n_i \to \min . \qquad \textbf{(12.9)}$$

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:

$$\begin{cases} b_0 \sum\limits_{i=1}^{l} n_i + b_1 \sum\limits_{i=1}^{l} x_i n_i = \sum\limits_{i=1}^{l} \bar{y}_i n_i , \\ b_0 \sum\limits_{i=1}^{l} x_i n_i + b_1 \sum\limits_{i=1}^{l} x_i^2 n_i = \sum\limits_{i=1}^{l} x_i \bar{y}_i n_i . \end{cases} \qquad \textbf{(12.10)}$$

Учитывая (**12.5**), преобразуем выражения:

$$\sum_{i=1}^{l}\bar{y}_i n_i=\sum_{i=1}^{l}\left(\frac{\sum_{j=1}^{m}y_j n_{ij}}{n_i}\right)n_i=\sum_{i=1}^{l}\sum_{j=1}^{m}y_j\,n_{ij}=\sum_{j=1}^{m}y_j\sum_{i=1}^{l}n_{ij}=\sum_{j=1}^{m}y_j\,n_j,$$

$$\sum_{i=1}^{l}x_i\bar{y}_i n_i=\sum_{i=1}^{l}x_i\left(\frac{\sum_{j=1}^{m}y_j\,n_{ij}}{n_i}\right)n_i=\sum_{i=1}^{l}\sum_{j=1}^{m}x_i y_j\,n_{ij}.$$

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:

$$\bar{x}=\frac{\sum_{i=1}^{l}x_i n_i}{n},\quad \bar{y}=\frac{\sum_{j=1}^{m}y_j n_j}{n}, \qquad (12.12)$$

$$\overline{xy}=\frac{\sum_{i=1}^{l}\sum_{j=1}^{m}x_i y_j n_{ij}}{n}, \qquad (12.13)$$

$$\overline{x^2}=\frac{\sum_{i=1}^{l}x_i^2 n_i}{n}. \qquad (12.14)$$

Подставляя значение $b_0=\bar{y}-b_1\bar{x}$ из первого уравнения системы (**12.11**) в уравнение регрессии (**12.8**), получим $y_x=\bar{y}-b_1\bar{x}+b_1x$, или

$$y_x-\bar{y}=b_1(x-\bar{x}). \qquad (12.15)$$

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:

$$y_x-\bar{y}=b_{yx}(x-\bar{x}). \qquad (12.16)$$

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

$$b_{yx} = b_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{s_x^2} = \frac{\mu}{s_x^2}, \qquad (12.17)$$

где s_x^2 — выборочная дисперсия переменной X (см. (**8.10**)):

$$s_x^2 = \overline{x^2} - \bar{x}^2 = \frac{\sum_{i=1}^{l} x_i^2 n_i}{n} - (x)^2; \qquad (12.18)$$

μ — выборочный корреляционный момент или выборочная ковариация[1]:

$$\mu = \overline{xy} - \bar{x}\,\bar{y} = \frac{\sum_{i=1}^{l}\sum_{j=1}^{m} x_i y_j n_{ij}}{n} - \bar{x}\,\bar{y}. \qquad (12.19)$$

Рассуждая аналогично и полагая уравнение регрессии (**12.4**) линейным, можно привести его к виду:

$$x_y - \bar{x} = b_{xy}(y - \bar{y}), \qquad (12.20)$$

где

$$b_{xy} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{y^2} - \bar{y}^2} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{s_y^2} = \frac{\mu}{s_y^2} \qquad (12.21)$$

— *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*;

$$s_y^2 = \overline{y^2} - \bar{y}^2 = \frac{\sum_{j=1}^{m} y_j^2 n_j}{n} - (\bar{y})^2 \qquad (12.22)$$

—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**)

$$\bar{x} = \frac{\sum_{i=1}^{l} u_i n_i}{n} \cdot k + c, \qquad (12.23)$$

$$\bar{y} = \frac{\sum_{j=1}^{m} v_j n_j}{n} \cdot k' + c', \qquad (12.24)$$

$$s_x^2 = \frac{\sum_{i=1}^{l} u_i^2 n_i}{n} \cdot k^2 - (\bar{x} - c)^2, \qquad (12.25)$$

$$s_y^2 = \frac{\sum_{j=1}^{m} v_j^2 n_j}{n} \cdot k'^2 - (\bar{y} - c')^2. \qquad (12.26)$$

Покажем, что в этом случае формула для ковариации μ (**12.19**) примет вид:

$$\mu = \frac{\sum_{i=1}^{l}\sum_{j=1}^{m} u_i v_j n_{ij}}{n} \cdot k \cdot k' - (\bar{x} - c)(\bar{y} - c'). \qquad (12.27)$$

□ Представим правую часть равенства (**12.27**) в виде:

$\overline{uv}kk' - (\bar{x} - c)(\bar{y} - c')$, 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,

$\overline{uv} = \frac{1}{kk'}\left(\overline{xy} - c'\overline{x} - c\overline{y} + cc'\right)$, откуда

$$\overline{uv}kk' - (\overline{x} - c)(\overline{y} - c') = \overline{xy} - c'\overline{x} - c\overline{y} + cc' - (\overline{x} - c)(\overline{y} - c') =$$

$$= \overline{xy} - \overline{x}\,\overline{y} = \mu \text{ (по определению (12.13))}.$$

▷ **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_i		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:

$$\frac{y_x - \bar{y}}{s_y} = \left(b_{yx} \frac{s_x}{s_y} \right) \frac{x - \bar{x}}{s_x}. \qquad (12.28)$$

В этой системе величина

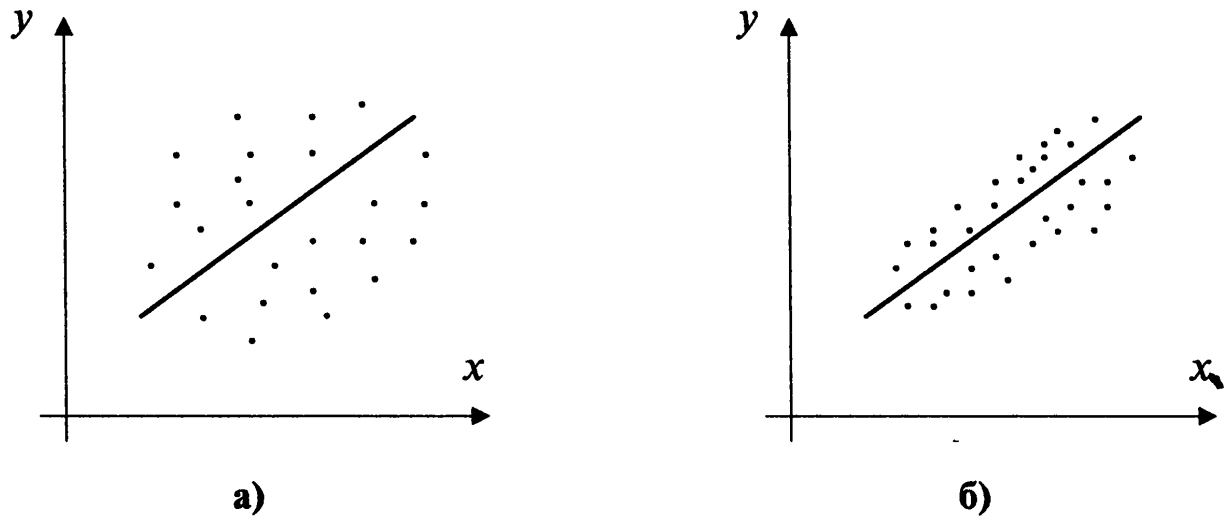
$$r = b_{yx} \frac{s_x}{s_y} \qquad (12.29)$$

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).



а) 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**):

$$r = \frac{\sum_{i=1}^{l}\sum_{j=1}^{m}(x_i - \bar{x})(y_j - \bar{y})n_{ij}}{ns_x s_y}; \qquad (12.34)$$

[1] См. ниже свойство **1** коэффициента корреляции.

$$r = \frac{n\sum_{i=1}^{l}\sum_{j=1}^{m} x_i y_j n_{ij} - \left(\sum_{i=1}^{l} x_i n_i\right)\left(\sum_{j=1}^{m} y_j n_j\right)}{\sqrt{n\sum_{i=1}^{l} x_i^2 n_i - \left(\sum_{i=1}^{l} x_i n_i\right)^2} \cdot \sqrt{n\sum_{j=1}^{m} y_j^2 n_j - \left(\sum_{j=1}^{m} y_j n_j\right)^2}}. \qquad \textbf{(12.35)}$$

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:

$$\sum_{i=1}^{14} x_i = 2{,}8 + 2{,}2 + \ldots + 6{,}0 + 9{,}0 = 64{,}2;$$

$$\sum_{i=1}^{14} x_i^2 = 2{,}8^2 + 2{,}2^2 + \ldots + 6{,}0^2 + 9{,}0^2 = 335{,}26;$$

$$\sum_{i=1}^{14} y_i = 6{,}7 + 6{,}9 + \ldots + 12{,}1 + 12{,}4 = 132{,}9;$$

$$\sum_{i=1}^{14} y_i^2 = 6{,}7^2 + 6{,}9^2 + \ldots + 12{,}1^2 + 12{,}4^2 = 1313{,}95;$$

$$\sum_{i=1}^{14} x_i y_i = 2{,}8 \cdot 6{,}7 + 2{,}2 \cdot 6{,}9 + \ldots + 6{,}0 \cdot 12{,}1 + 9{,}0 \cdot 12{,}4 = 650{,}99.$$

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

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum x_i\right)^2}\cdot\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} = \tag{12.35'}$$

$$= \frac{14\cdot 650{,}99 - 64{,}2\cdot 132{,}9}{\sqrt{14\cdot 335{,}26 - 64{,}2^2}\sqrt{14\cdot 1313{,}95 - 132{,}4^2}} = 0{,}898,$$

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.

$$-1 \le r \le 1. \tag{12.36}$$

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* $r = \pm 1$ *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.*

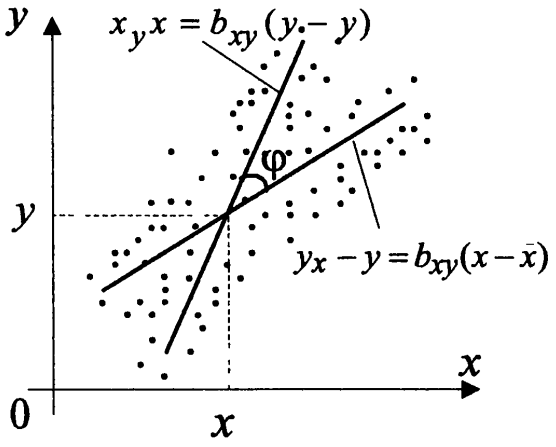


Рис. 12.3

□ 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.

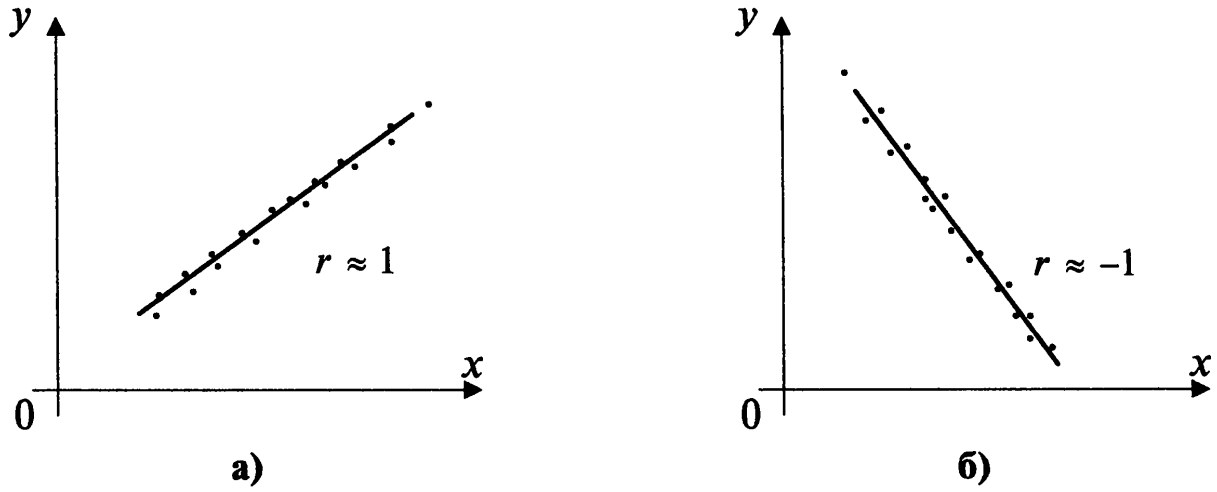
$k_2 = \dfrac{1}{b_{xy}}$, используя соответствующую формулу аналитической геометрии:

$$\operatorname{tg}\varphi = \frac{k_2 - k_1}{1 + k_2 k_1} = \frac{1 - b_{yx} b_{xy}}{b_{xy} + b_{yx}},$$

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 (у́же 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 $r = 0$ л и н е й н а я 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). ■

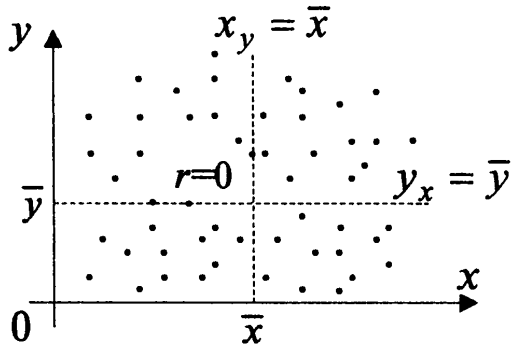


Рис. 12.5

Равенство $r = 0$ говорит лишь об отсутствии л и н е й н о й 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.

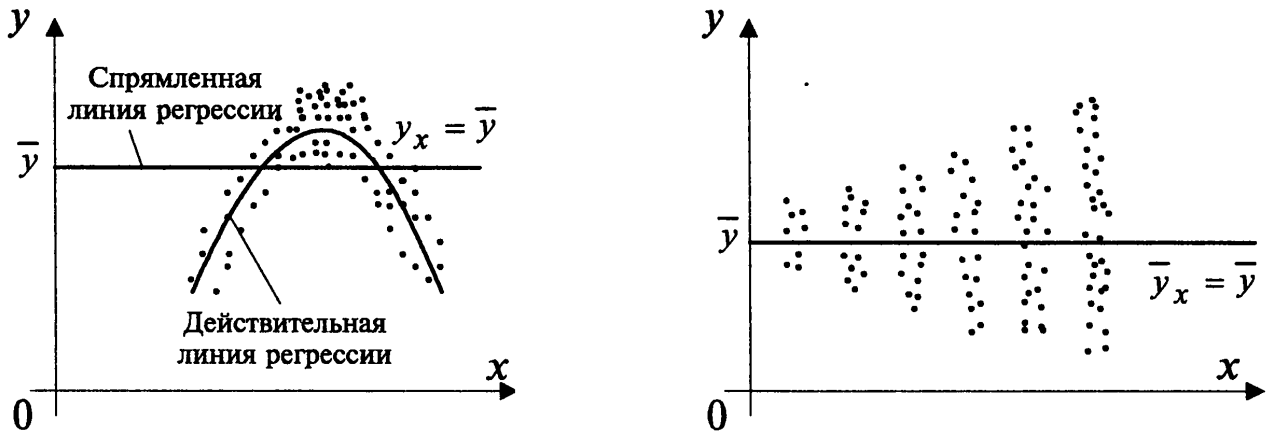


Рис. 12.6

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 ковариация* 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 коэффици-

ент 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

$$\eta_{yx} = \sqrt{\frac{\delta_{iy}^2}{s_y^2}} \qquad \textbf{(12.58)}$$

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*:

$$\eta_{yx} = \sqrt{\frac{\delta_{ix}^2}{s_x^2}}. \qquad \textbf{(12.59)}$$

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. $\eta_{yx} \neq \eta_{xy}$, 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*:

$$R_{yx}=\sqrt{\frac{\delta_y^2}{s_y^2}}=\sqrt{1-\frac{s_y'^2}{s_y^2}}, \qquad \textbf{(12.60)}$$

где дисперсии δ_y^2 и $s_y'^2$ определяются по формулам **(12.54)**—**(12.56)**, в которых групповые средние $\bar{y}_i$ заменены условными средними y_{x_i}, вычисленными по уравнению регрессии **(12.16)**.

Подобно R_{yx} вводится и *индекс корреляции X по Y*:

$$R_{xy}=\sqrt{\frac{\delta_x^2}{s_x^2}}=\sqrt{1-\frac{s_x'^2}{s_x^2}}. \qquad \textbf{(12.61)}$$

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:

$$0 \le |r| \le R \le \eta \le 1. \qquad \textbf{(12.62)}$$

Покажем, что в случае линейной 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 $n_i=1$ ($i=1,2,\ldots,l$).

По формуле **(12.60)**

$$R_{yx}=\sqrt{\frac{\delta_y^2}{s_y^2}}=\sqrt{\frac{\sum\limits_{i=1}^{n}(y_{x_i}-\bar{y})^2}{\sum\limits_{i=1}^{n}(y_i-\bar{y})^2}}=\sqrt{\frac{\sum\limits_{i=1}^{n}b_{yx}^2(x_i-\bar{x})^2}{\sum\limits_{i=1}^{n}(y_i-\bar{y})^2}}=|b_{yx}|\sqrt{\frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2/n}{\sum\limits_{i=1}^{n}(y_i-\bar{y})^2/n}}$$

(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 л и н е й н о с т и 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**) $\eta_{yx} = \sqrt{\dfrac{10{,}36}{18{,}23}} = \sqrt{0{,}568} = 0{,}754$. 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 p_{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} ($i, j = 1,2, \ldots, p$) 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 $q \leq (p-2)$.

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* $\rho_{i\,12\ldots p}$, который является обобщением парного коэффициента корреляции ρ_{ij}. *Выборочный множественный*, или *совокупный*, *коэффициент корреляции* $R_{i.12\ldots p}$, являющийся оценкой $R_{i.12\ldots p}$, может быть вычислен по формуле:

$$R_{i.12\ldots p} = \sqrt{1 - \frac{|q_p|}{q_{ii}}}, \qquad \textbf{(12.67)}$$

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

$$r_{ij.12\,..p} = \frac{-q_{ij}}{\sqrt{q_{ii}q_{jj}}}, \qquad \textbf{(12.70)}$$

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

$$r_{ij\,k} = \frac{r_{ij} - r_{ik}\cdot r_{jk}}{\sqrt{\left(1-r_{ik}^2\right)\left(1-r_{jk}^2\right)}}. \qquad \textbf{(12.71)}$$

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:

$$r_{12.3} = \frac{0{,}20^2 - 0{,}41 \cdot 0{,}82}{\sqrt{\left(1-0{,}41^2\right)\left(1-0{,}82^2\right)}} = -0{,}26$$

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

$$F = \frac{\left(\eta^2 - r^2\right)(n-m)}{(m-2)\left(1-\eta^2\right)} \qquad \textbf{(12.72)}$$

(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 порядок, 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 ран-

жированы 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 $s_\tau = \sqrt{\dfrac{2(2n+5)}{9n(n-1)}}$. 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]:

$$W = \frac{12\sum_{i=1}^{n} D^2}{m^2(n^3 - n)}, \qquad \textbf{(12.79)}$$

где n — число объектов,

m — число анализируемых порядковых переменных,

$$D = \sum_{j=1}^{m} r_{ij} - \frac{m(n+1)}{2} \qquad \textbf{(12.80)}$$

— 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

$$m(n-1)W > \chi^2_{\alpha,n-1}. \qquad \textbf{(12.81)}$$

▷ **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**) $W = \frac{12 \cdot 416}{5^2\left(7^3 - 7\right)} = 0{,}594$. 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: $\bar{x}=16{,}2$ (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 $\alpha = 0{,}05$.